Manifold Sampling for Composite Nonconvex Nonsmooth Optimization

Kamil Khan, Jeffrey Larson, Matt Menickelly, Stefan Wild

Argonne National Laboratory

August 5, 2019

Problem setup

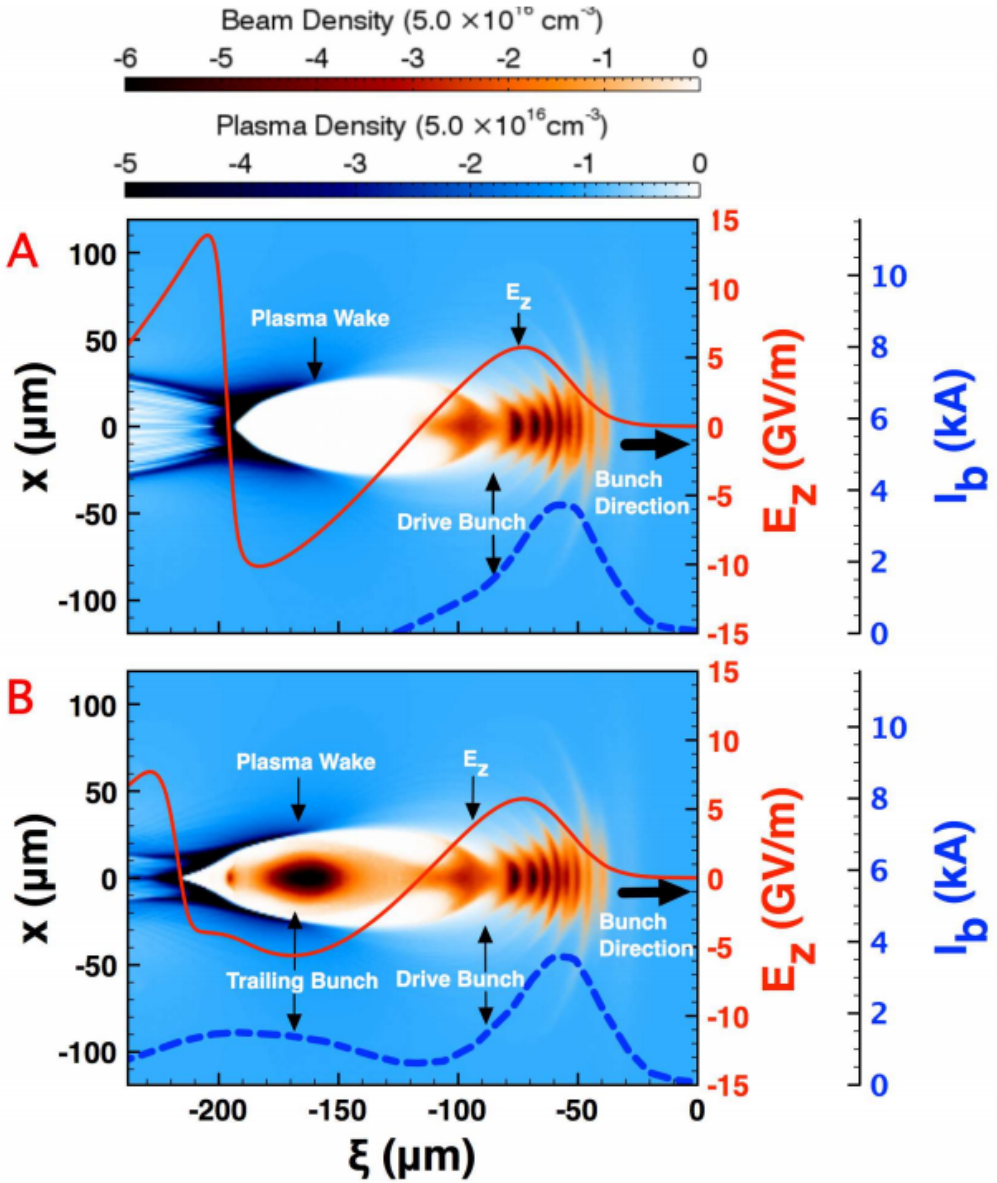
- Nonsmooth, composite optimization

$$\underset{x}{\text{minimize}}\, f(x) = h(F(x))$$

  nonsmooth $h\colon \mathbb{R}^p \to \mathbb{R}$ (with a known structure), smooth $F\colon \mathbb{R}^n \to \mathbb{R}^p$ (expensive to evaluate).
- Idea: Build p models, one for each component of F. Use model gradients in place of ∇F.
- Requires a *manifold representation* of h.
- Example: censored loss:

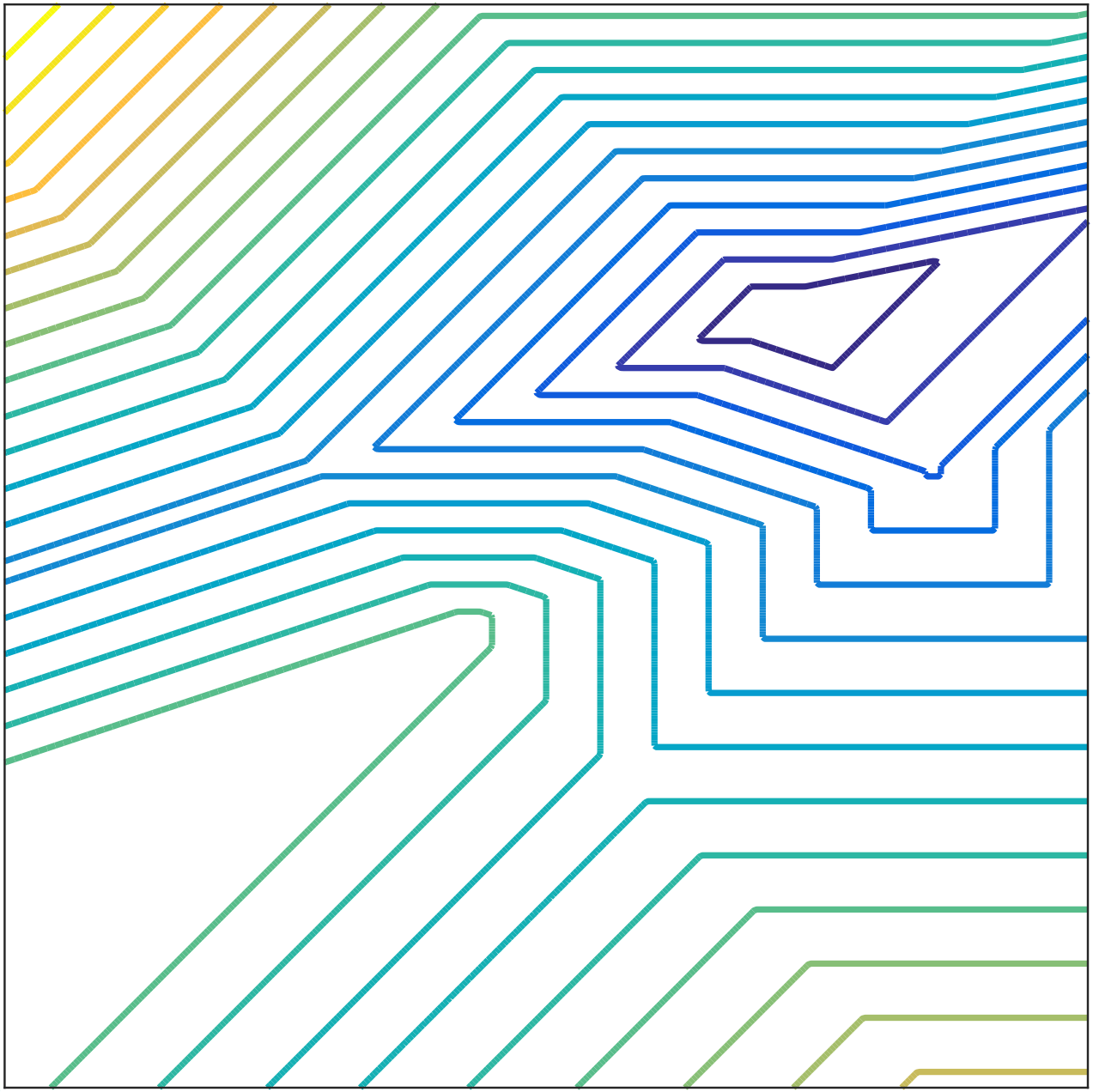
$$f(x) = \sum_{i=1}^{p} |d_i - \max\{c_i, F_i(x)\}|$$

Computers/Simulations!

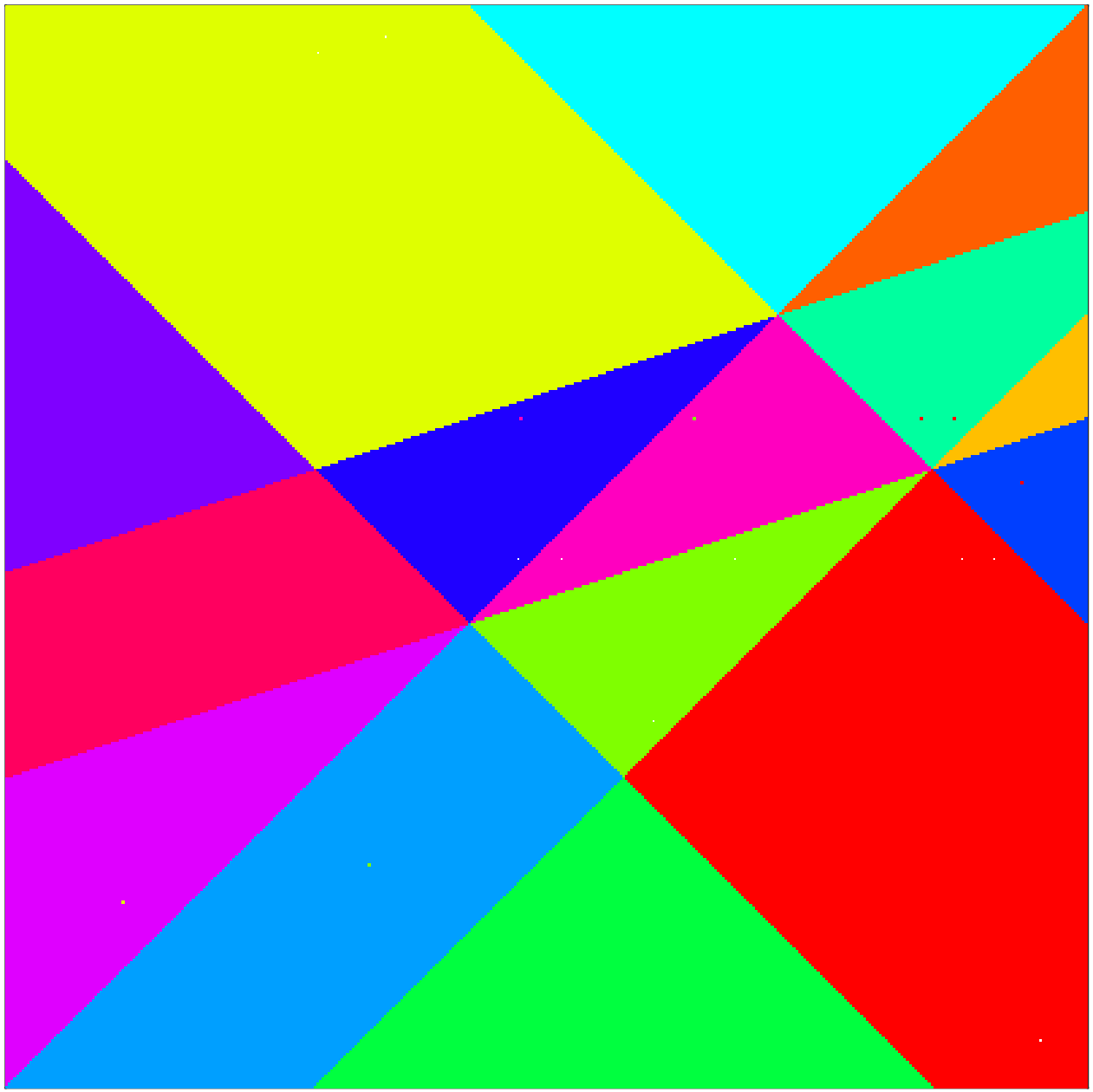


Computers/Simulations!

Censored ℓ_1 loss



Censored ℓ_1 loss



Manifold representation

- $h(y) = \max_{i \in \{1,\ldots,p\}} y_i$ p manifolds

Manifold representation

- $h(y) = \max_{i \in \{1,\dots,p\}} y_i$ — p manifolds
- $h(y) = \|y\|_\infty = \max_{i \in \{1,\dots,p\}} |y_i|$ — $2p$ manifolds

Manifold representation

- $h(y) = \max_{i \in \{1,\dots,p\}} y_i$ — p manifolds
- $h(y) = \|y\|_\infty = \max_{i \in \{1,\dots,p\}} |y_i|$ — $2p$ manifolds
- $h(y) = \max_{i \in I_1}\{y_i\} - \min_{i \in I_2}\{y_i\}$ — $|I_1|\,|I_2|$ manifolds

Manifold representation

- $h(y) = \max_{i \in \{1,\ldots,p\}} y_i$ — p manifolds
- $h(y) = \|y\|_\infty = \max_{i \in \{1,\ldots,p\}} |y_i|$ — $2p$ manifolds
- $h(y) = \max_{i \in I_1}\{y_i\} - \min_{i \in I_2}\{y_i\}$ — $|I_1|\,|I_2|$ manifolds
- $h(y) = \|y\|_1 = \sum_{i=1}^{p} |y_i|$ — 2^p manifolds

Manifold representation

- $h(y) = \max_{i \in \{1,\ldots,p\}} y_i$ — p manifolds
- $h(y) = \|y\|_\infty = \max_{i \in \{1,\ldots,p\}} |y_i|$ — $2p$ manifolds
- $h(y) = \max_{i \in I_1}\{y_i\} - \min_{i \in I_2}\{y_i\}$ — $|I_1|\,|I_2|$ manifolds
- $h(y) = \|y\|_1 = \sum_{i=1}^{p} |y_i|$ — 2^p manifolds
- $h(y) = \sum_{i=1}^{p} |d_i - \max\{c_i, y_i\}|$ — 3^p manifolds. If $p = 45$, approximately 3×10^{21} potential manifolds.

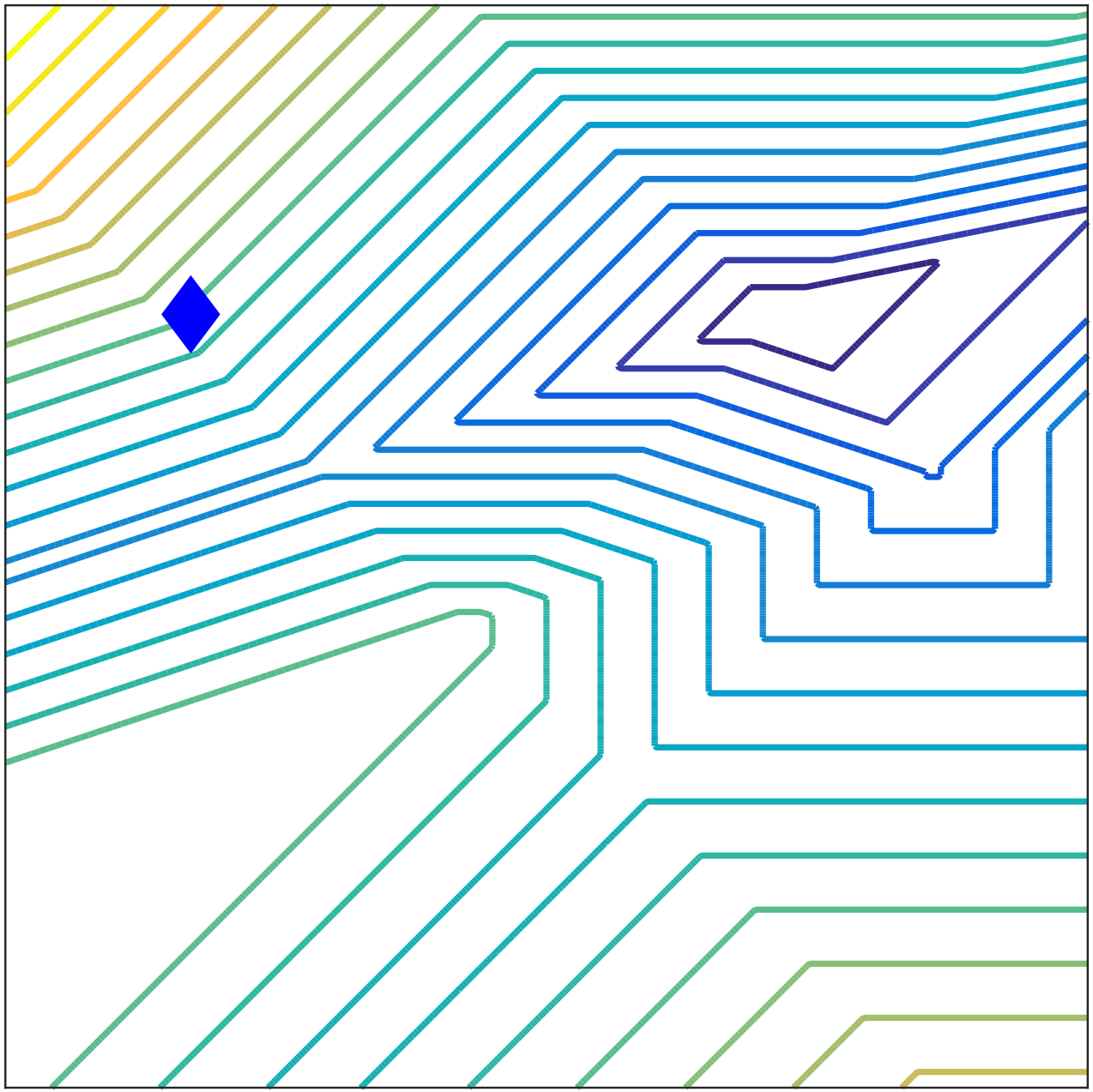
Manifold representation

- $h(y) = \max_{i \in \{1,\ldots,p\}} y_i$ — p manifolds
- $h(y) = \|y\|_\infty = \max_{i \in \{1,\ldots,p\}} |y_i|$ — $2p$ manifolds
- $h(y) = \max_{i \in I_1}\{y_i\} - \min_{i \in I_2}\{y_i\}$ — $|I_1|\,|I_2|$ manifolds
- $h(y) = \|y\|_1 = \sum_{i=1}^{p} |y_i|$ — 2^p manifolds
- $h(y) = \sum_{i=1}^{p} |d_i - \max\{c_i, y_i\}|$ — 3^p manifolds. If $p = 45$, approximately 3×10^{21} potential manifolds.

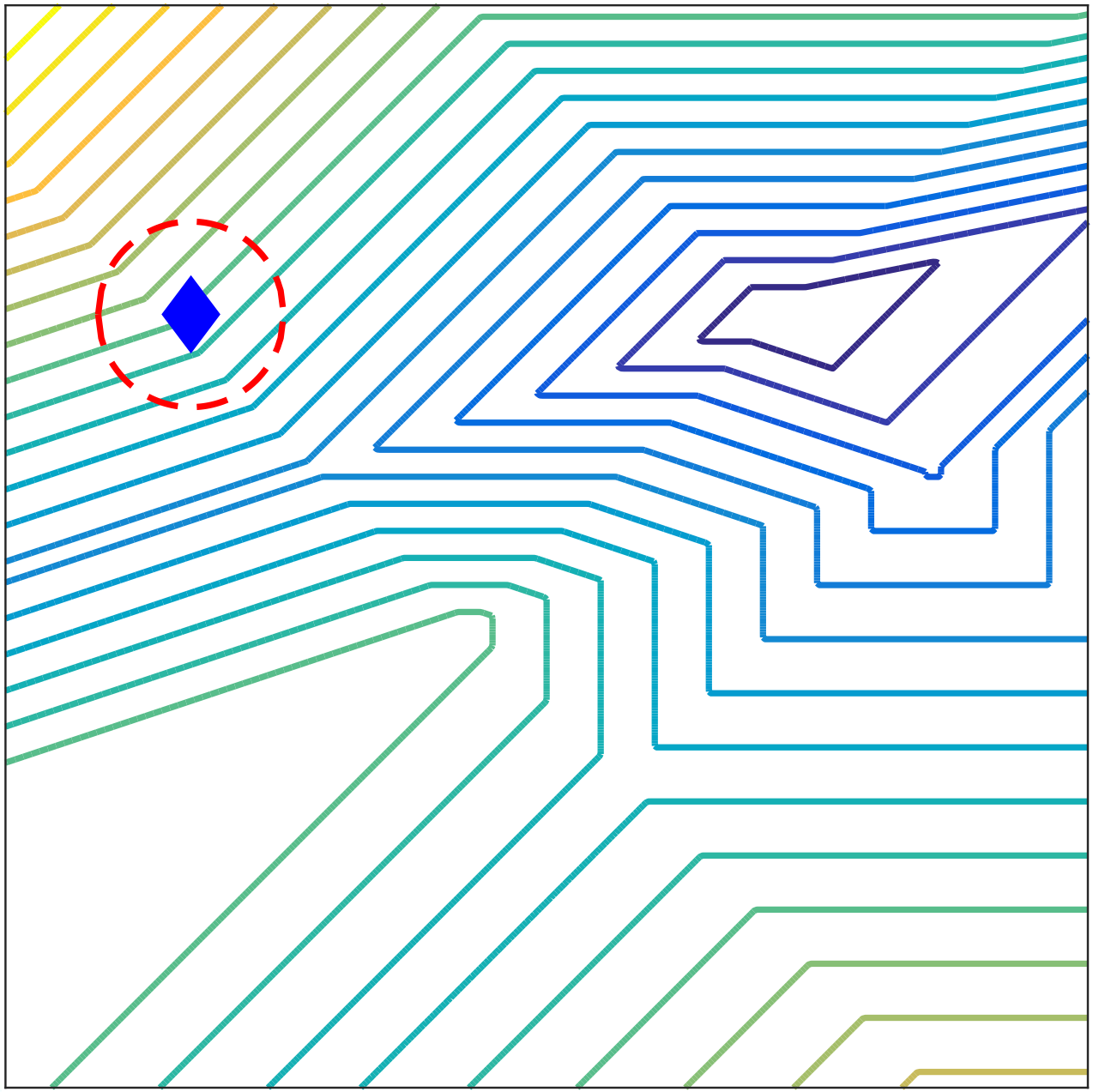
User scripts need to calculate:

$f(x)$, $F(x)$, $\mathbb{H}(F(x))$, $\{\nabla h_i(F(x)) : i \in \mathbb{H}(F(x))\}$, $\{h_i(F(x)) : i \in \mathbb{G}\}$,

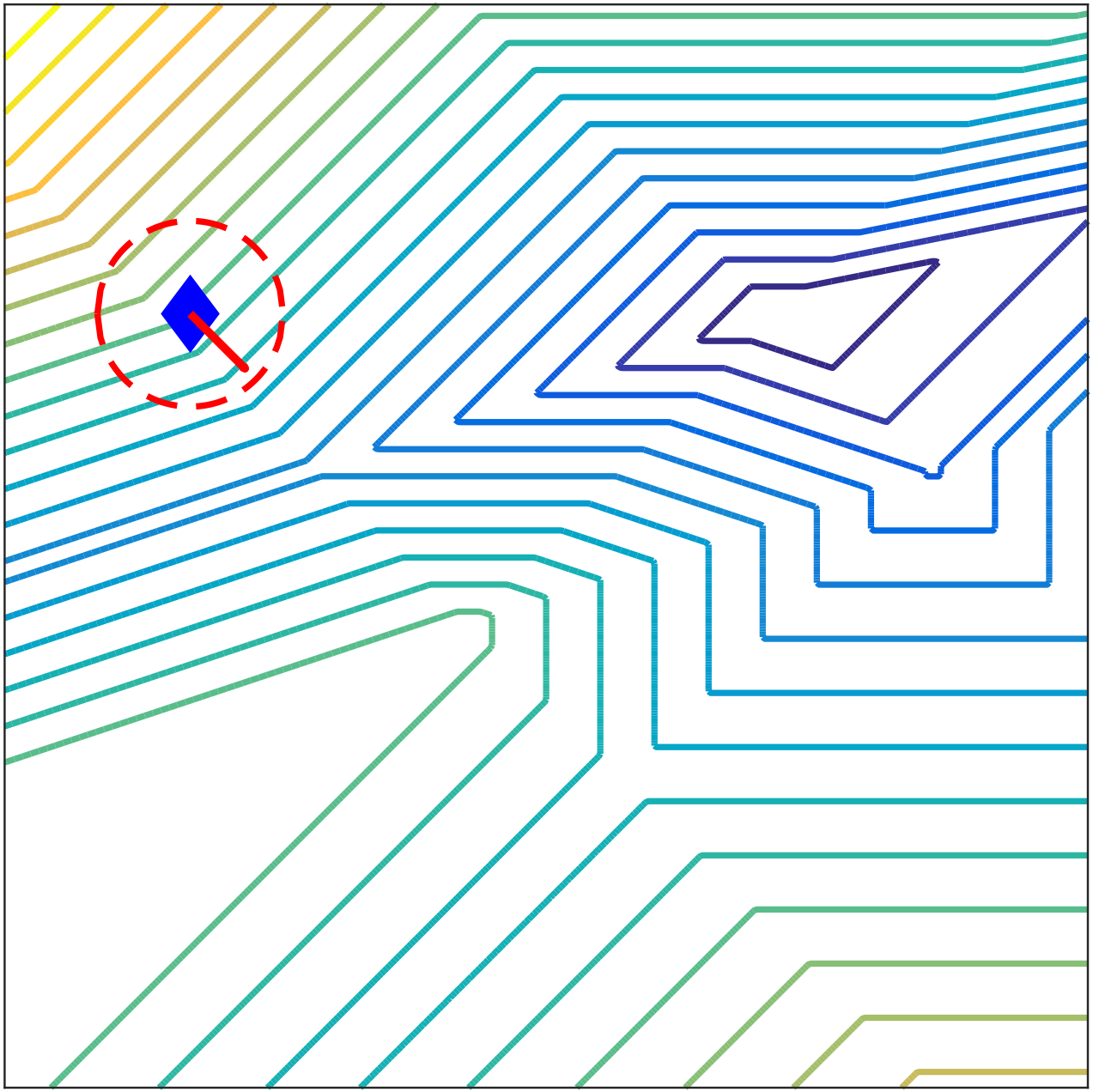
Manifold Sampling



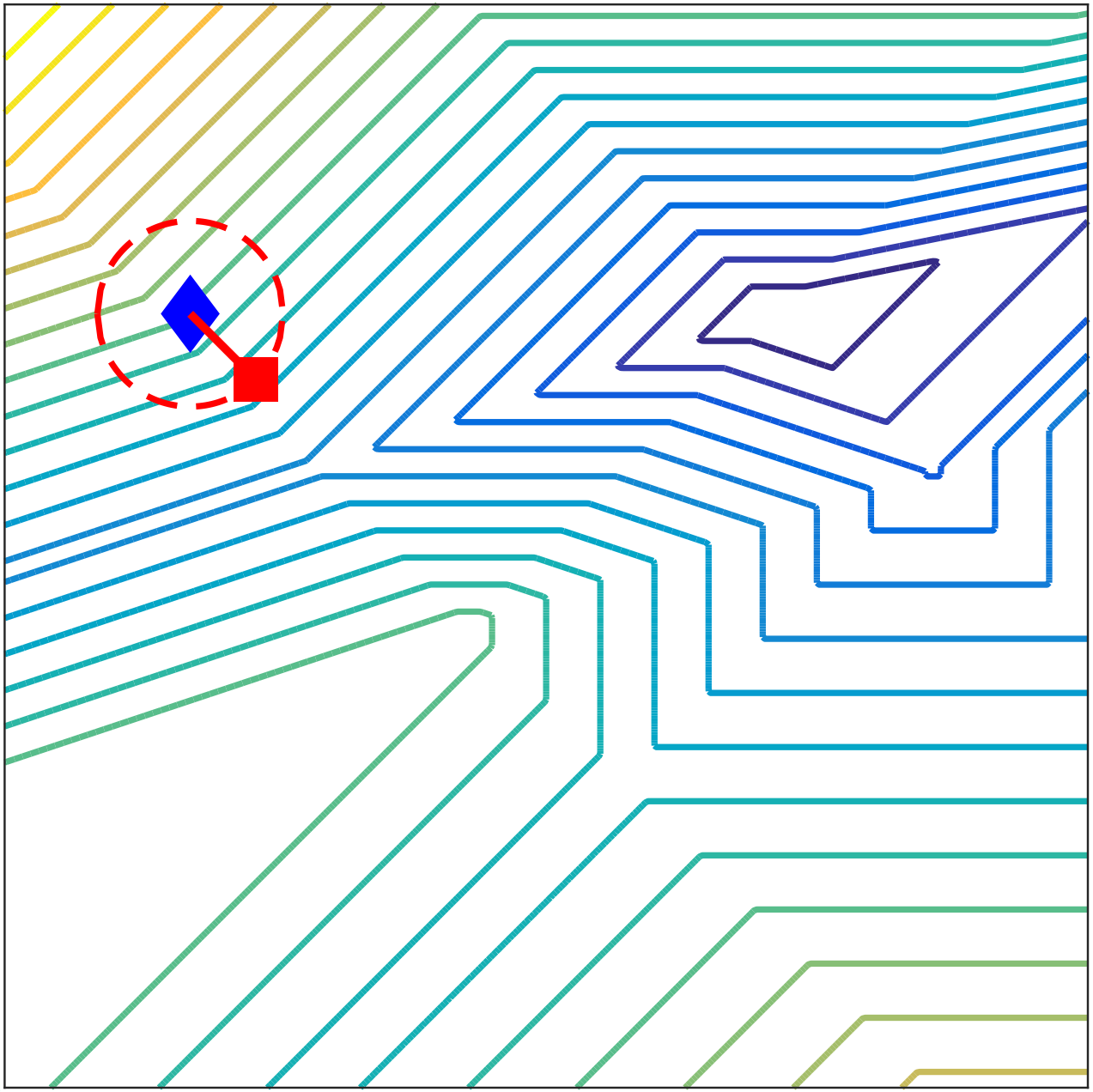
Manifold Sampling



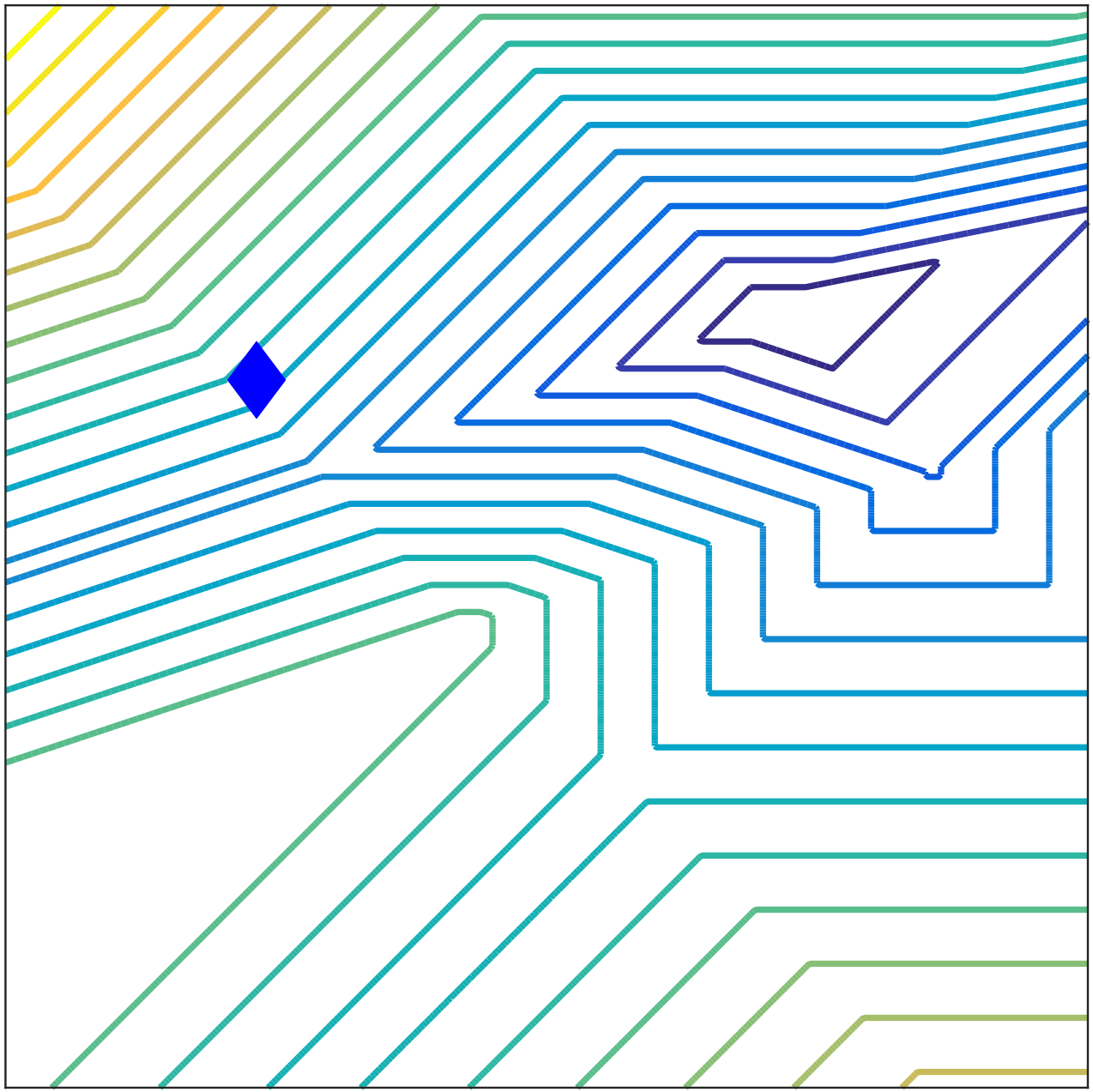
Manifold Sampling



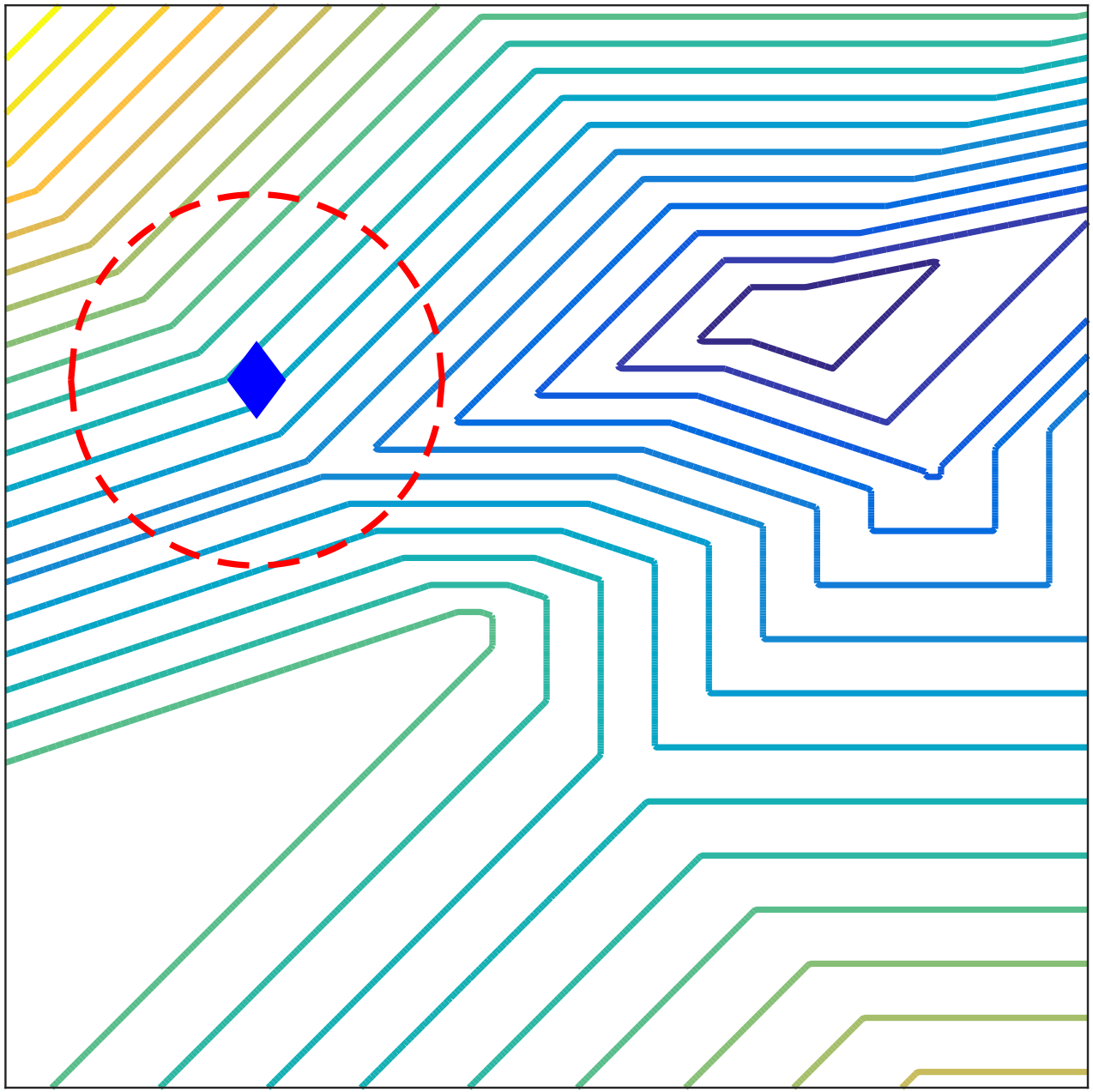
Manifold Sampling



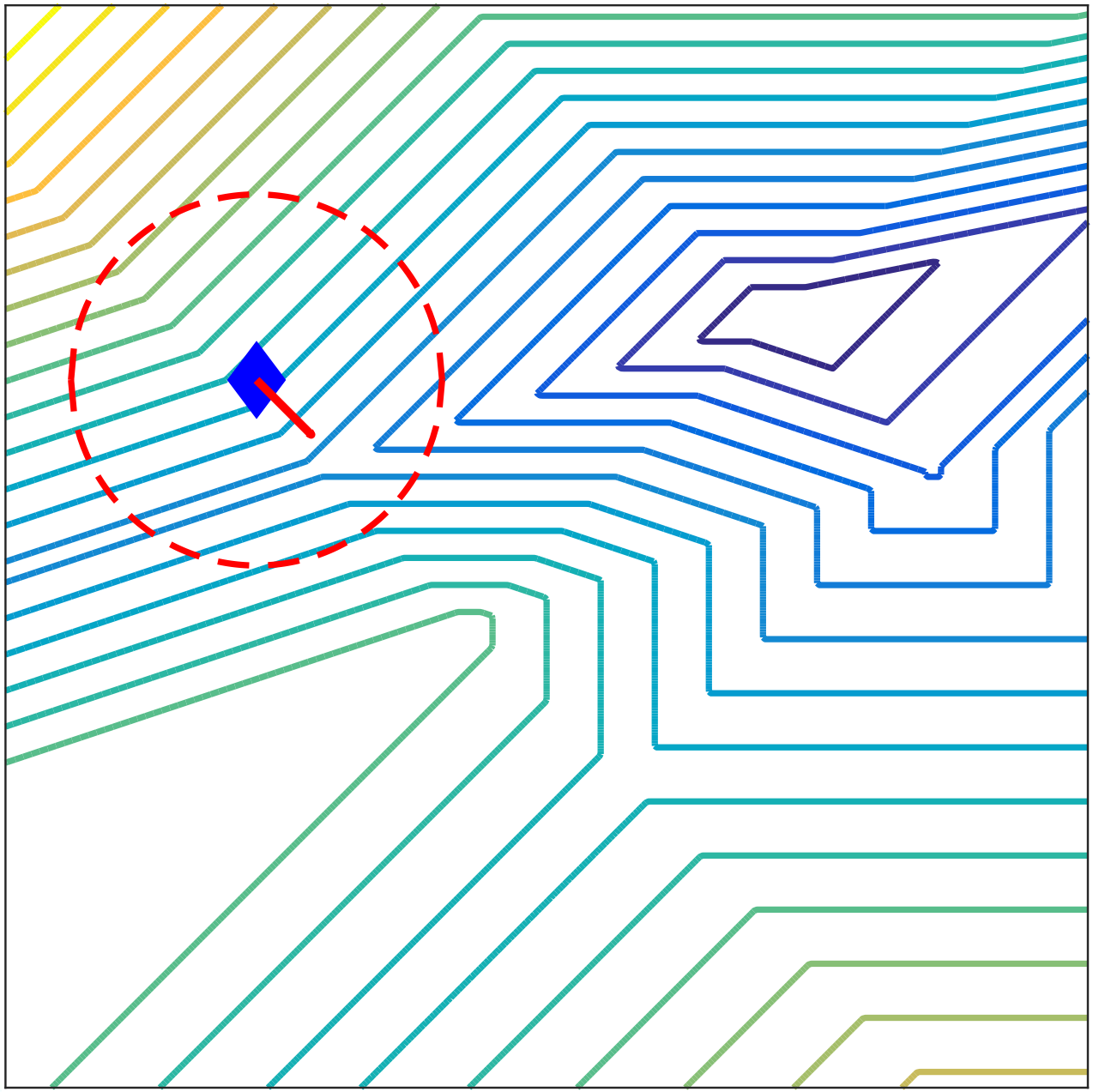
Manifold Sampling



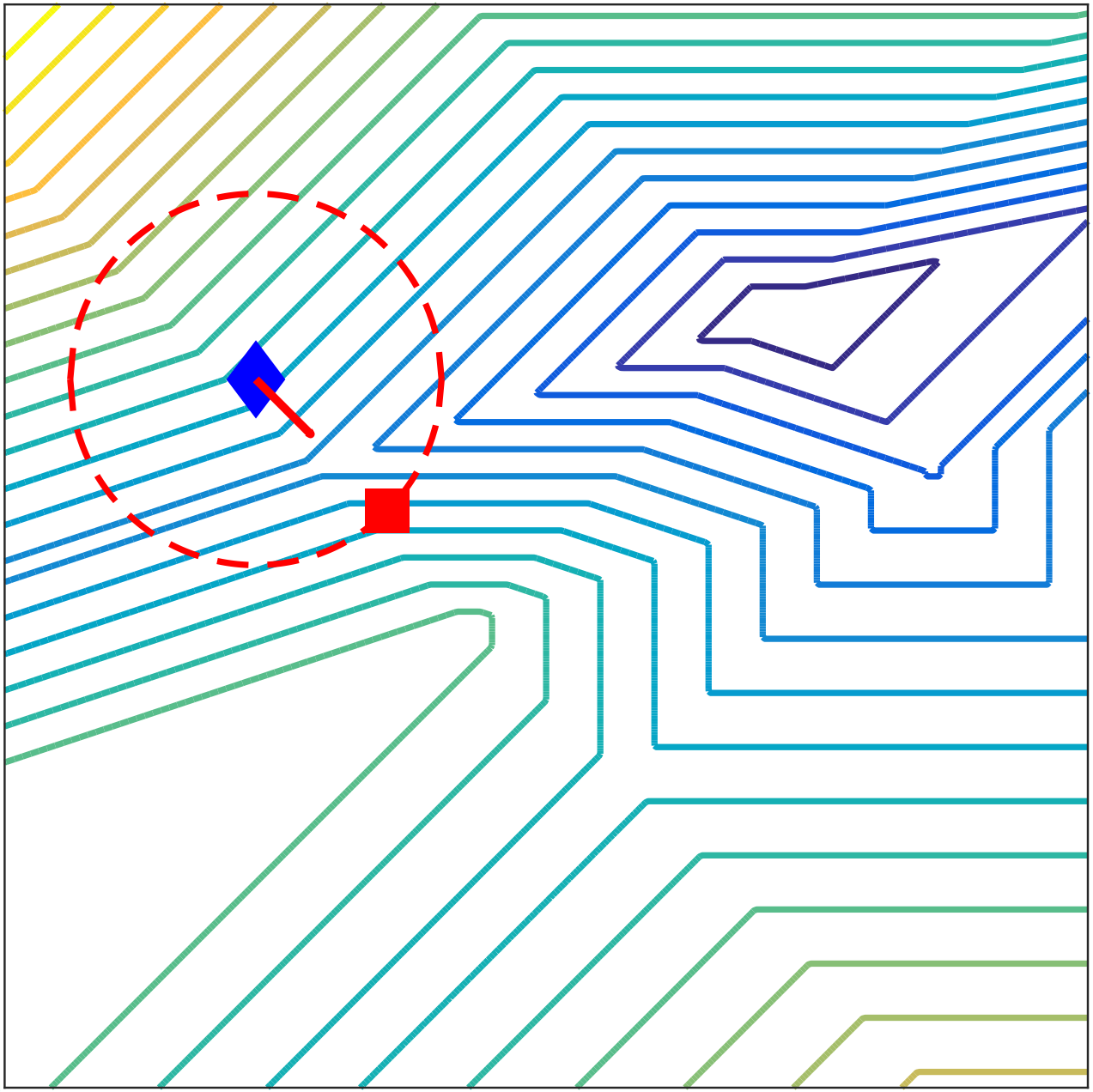
Manifold Sampling



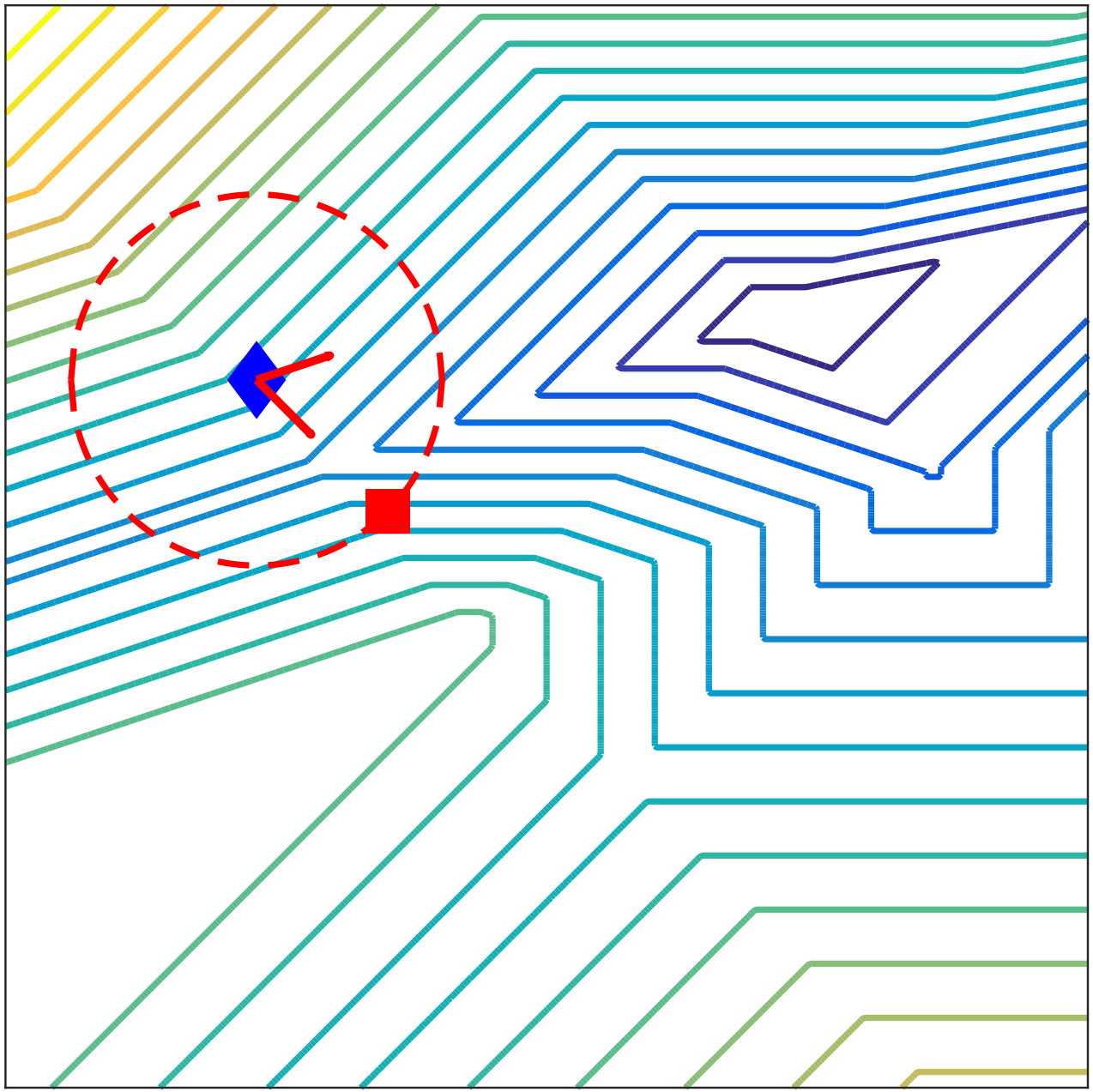
Manifold Sampling



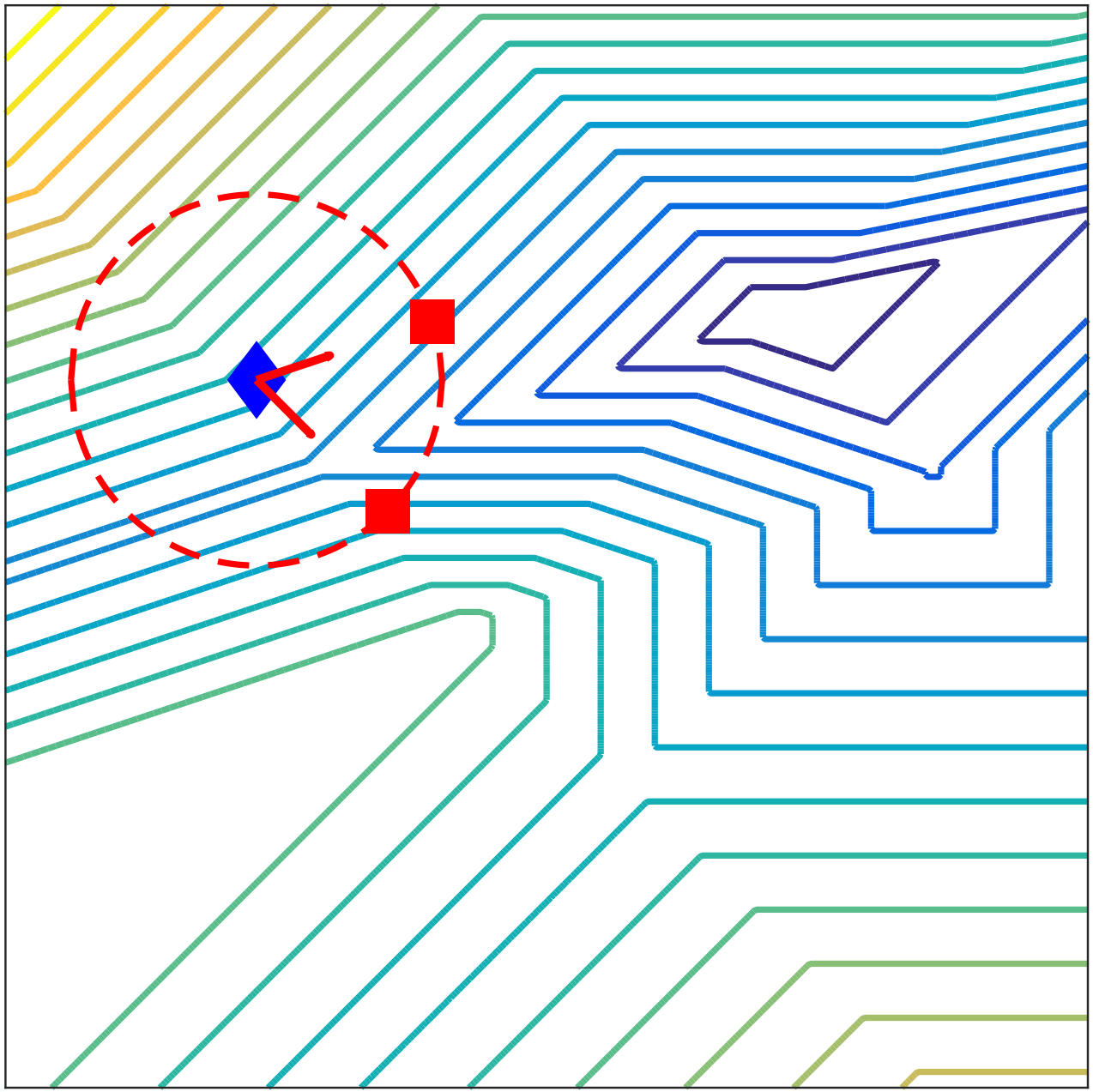
Manifold Sampling



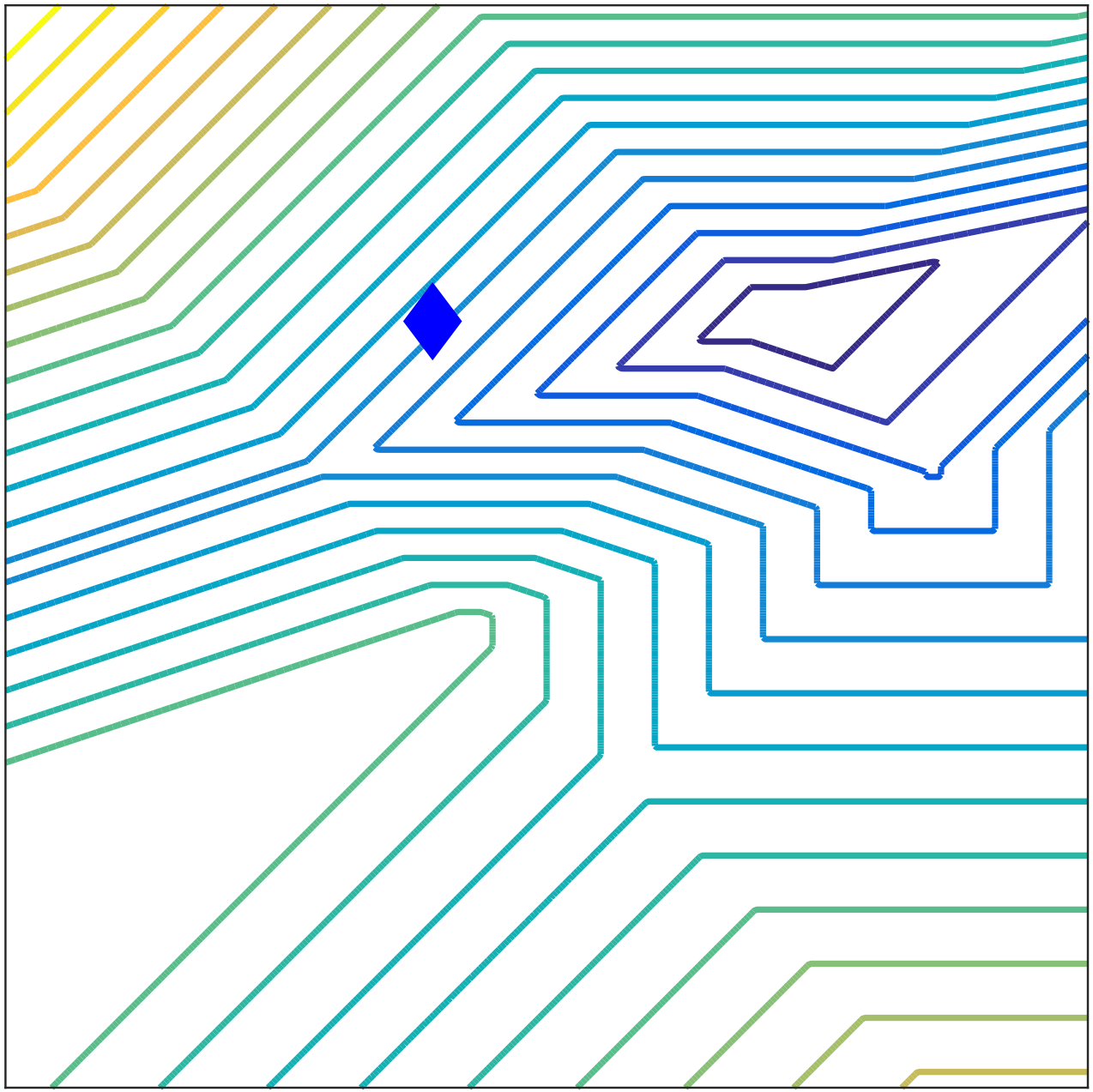
Manifold Sampling



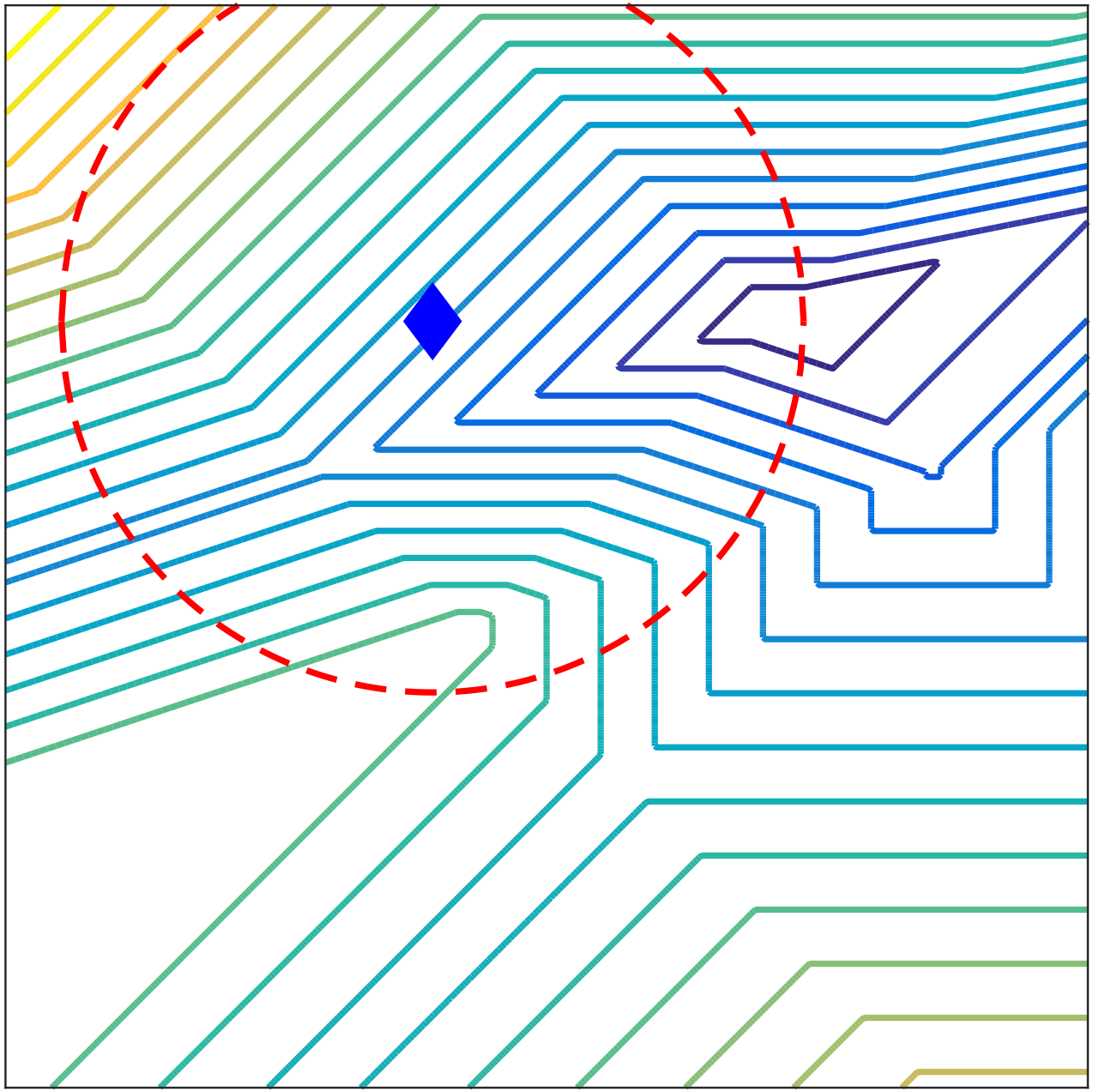
Manifold Sampling



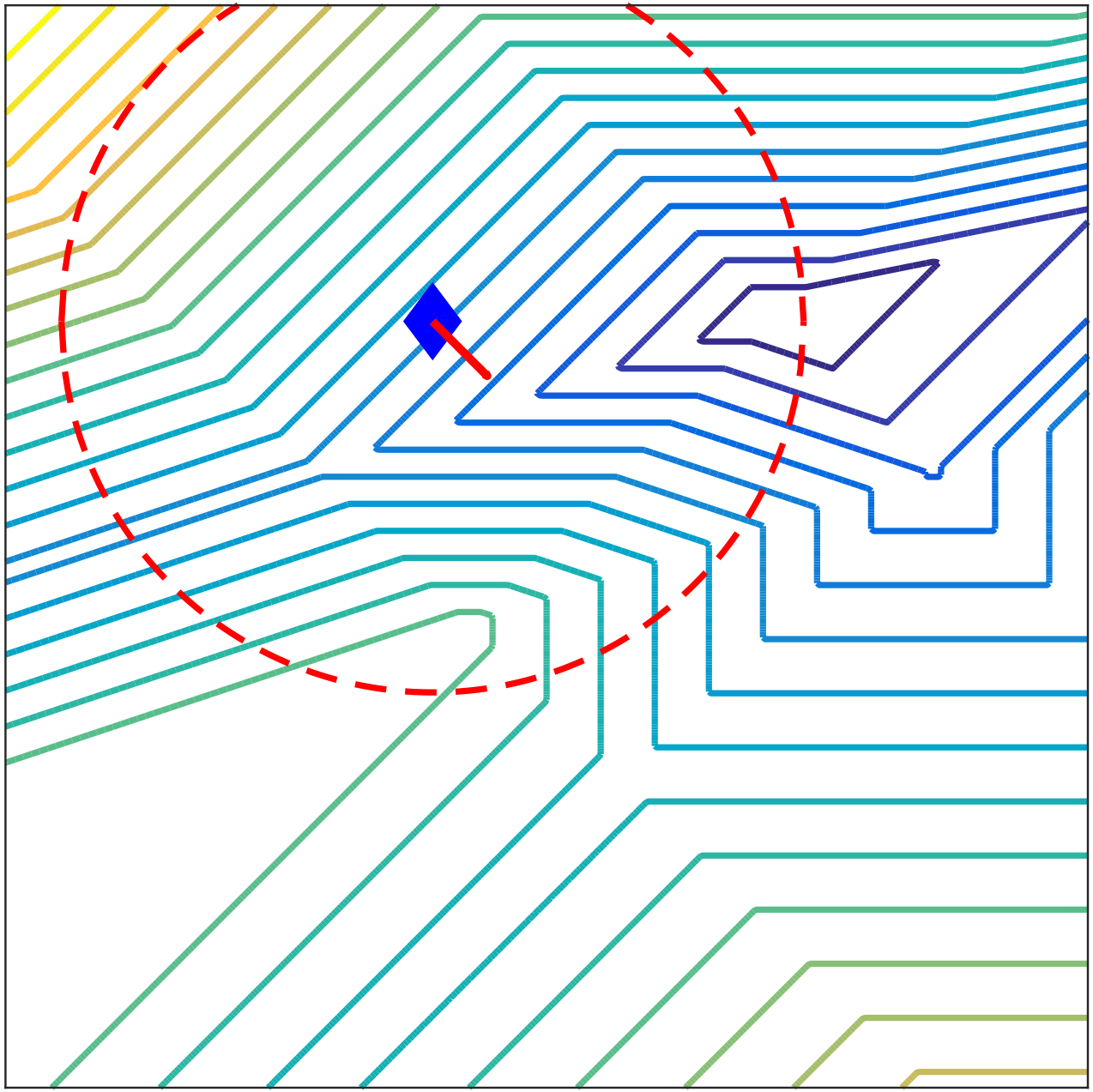
Manifold Sampling



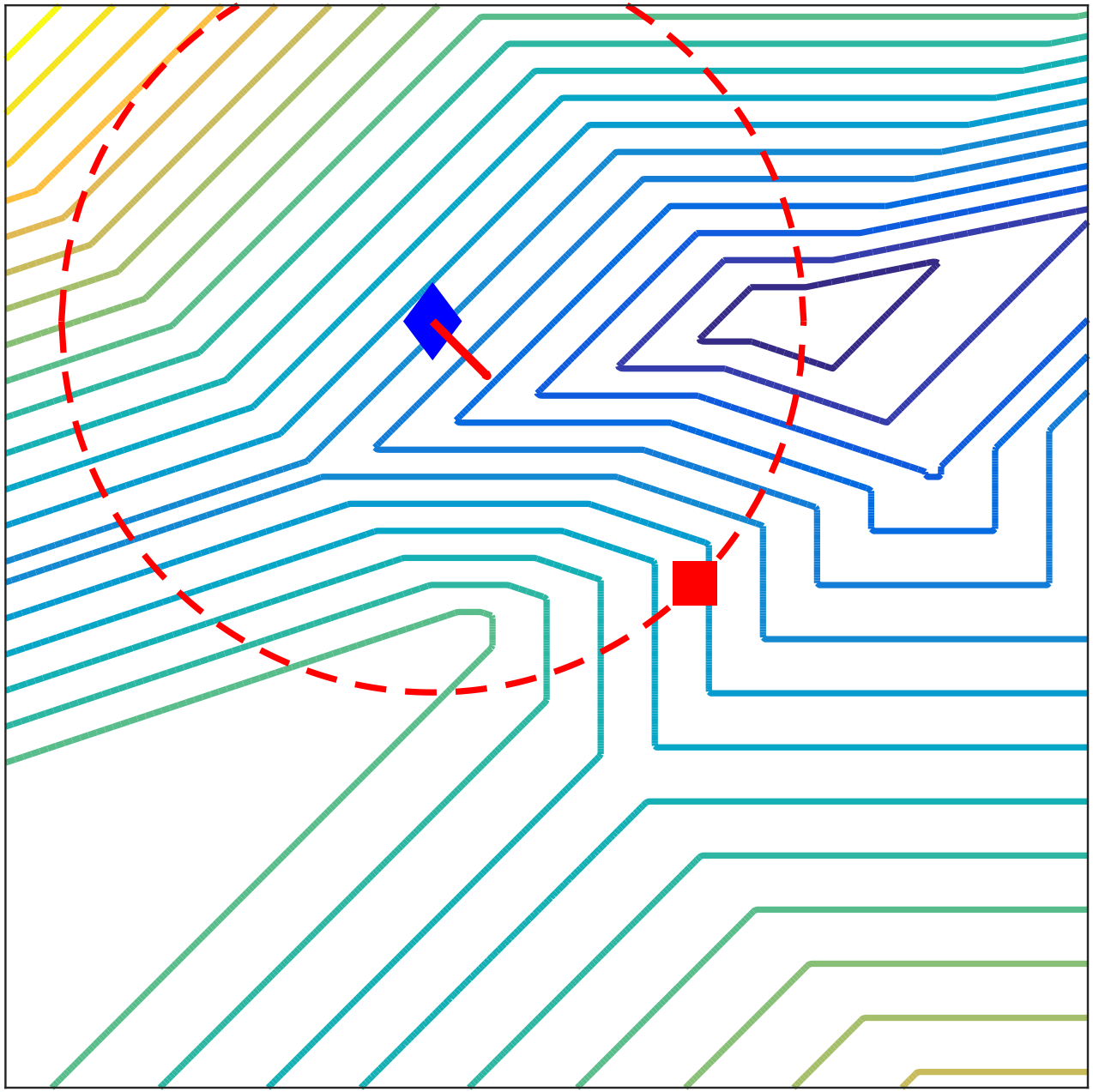
Manifold Sampling



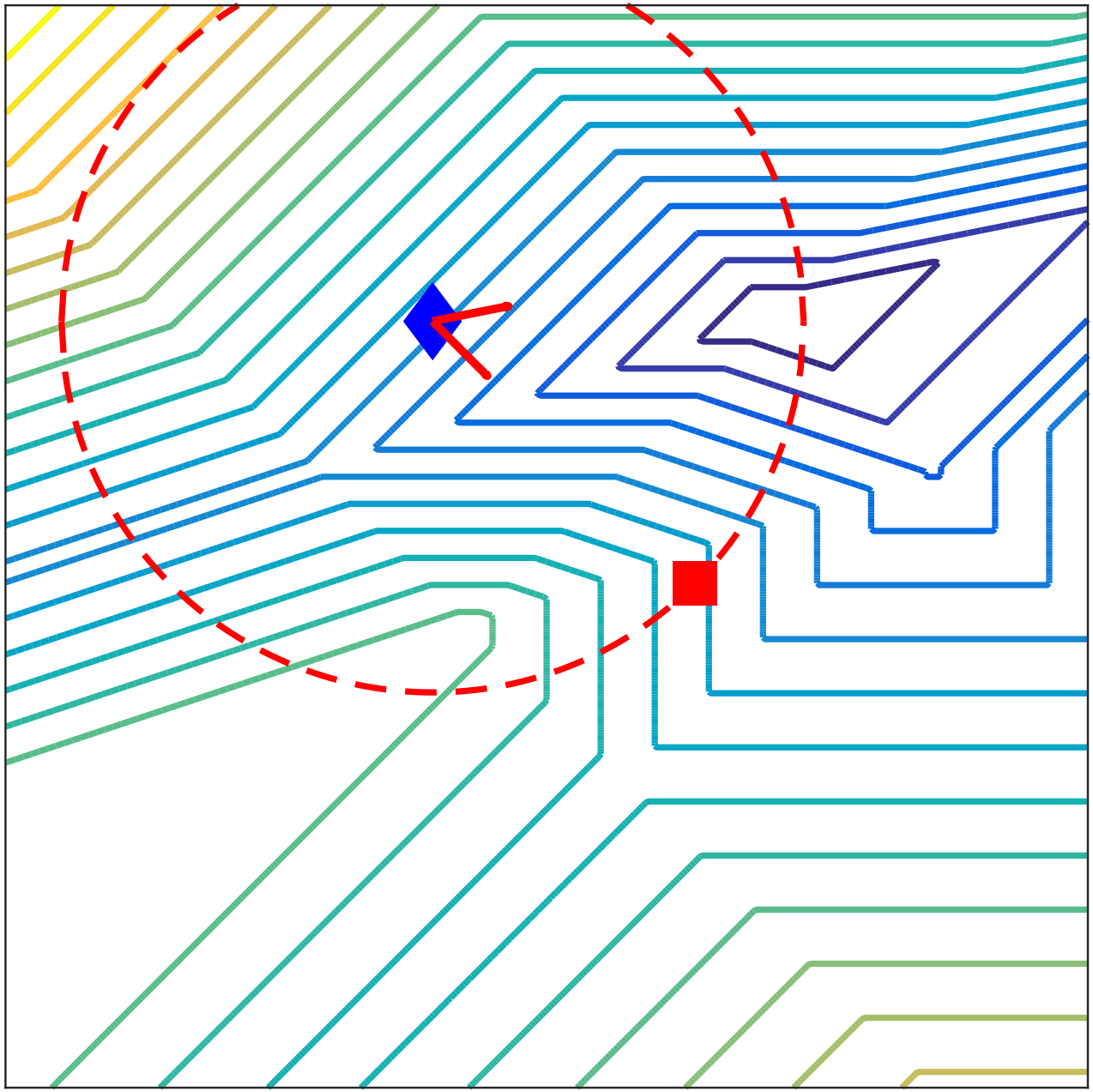
Manifold Sampling



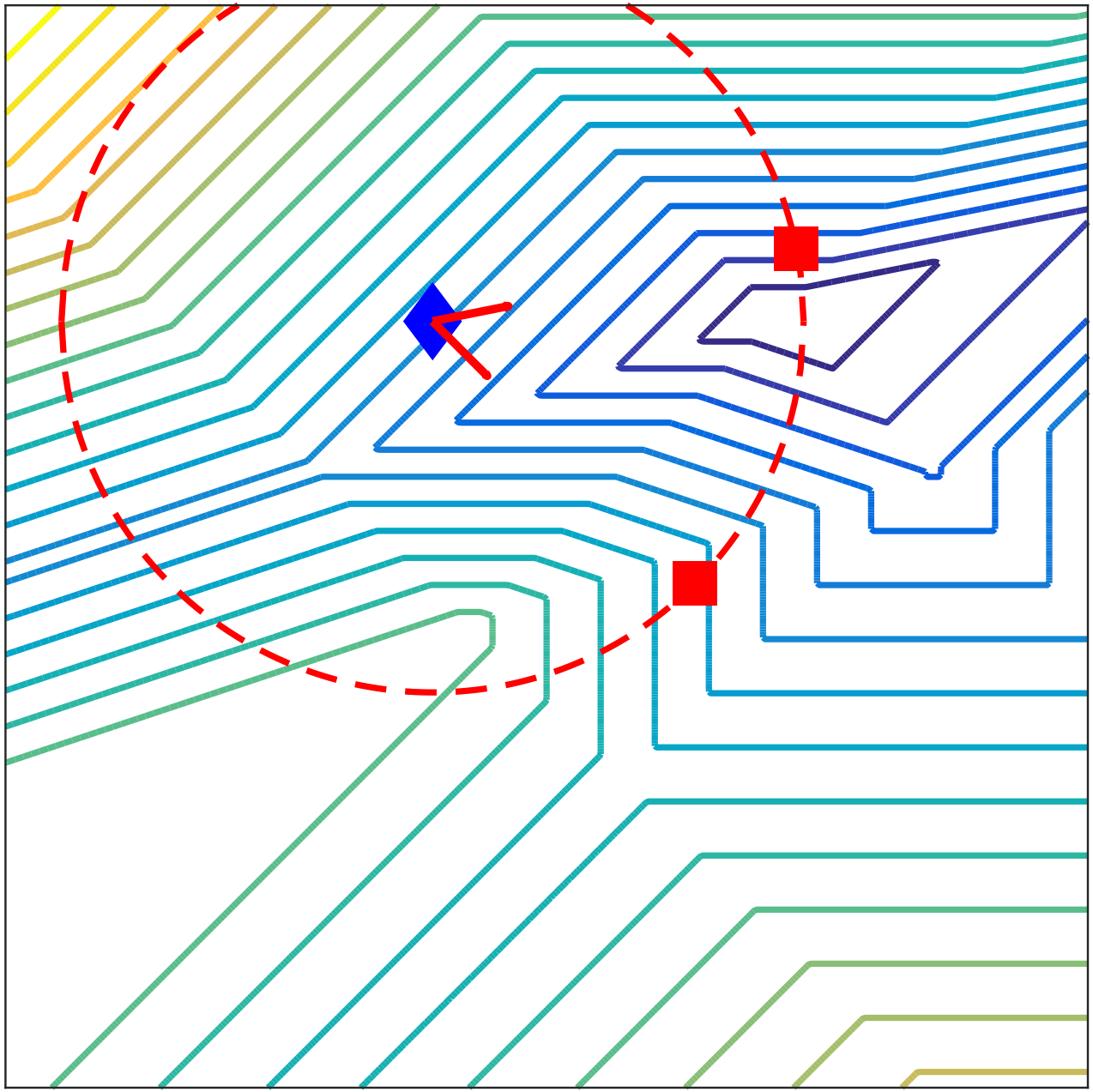
Manifold Sampling



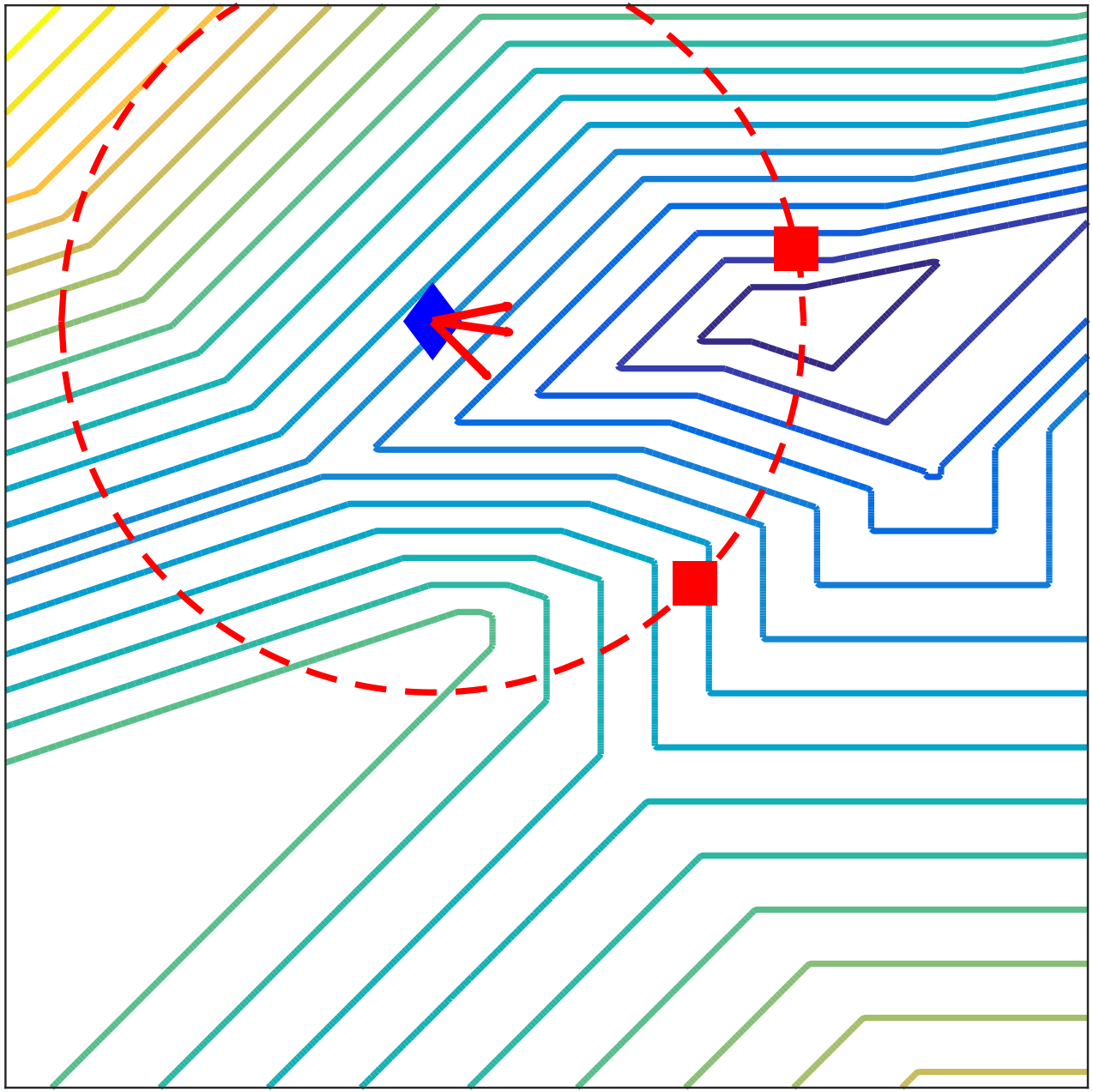
Manifold Sampling



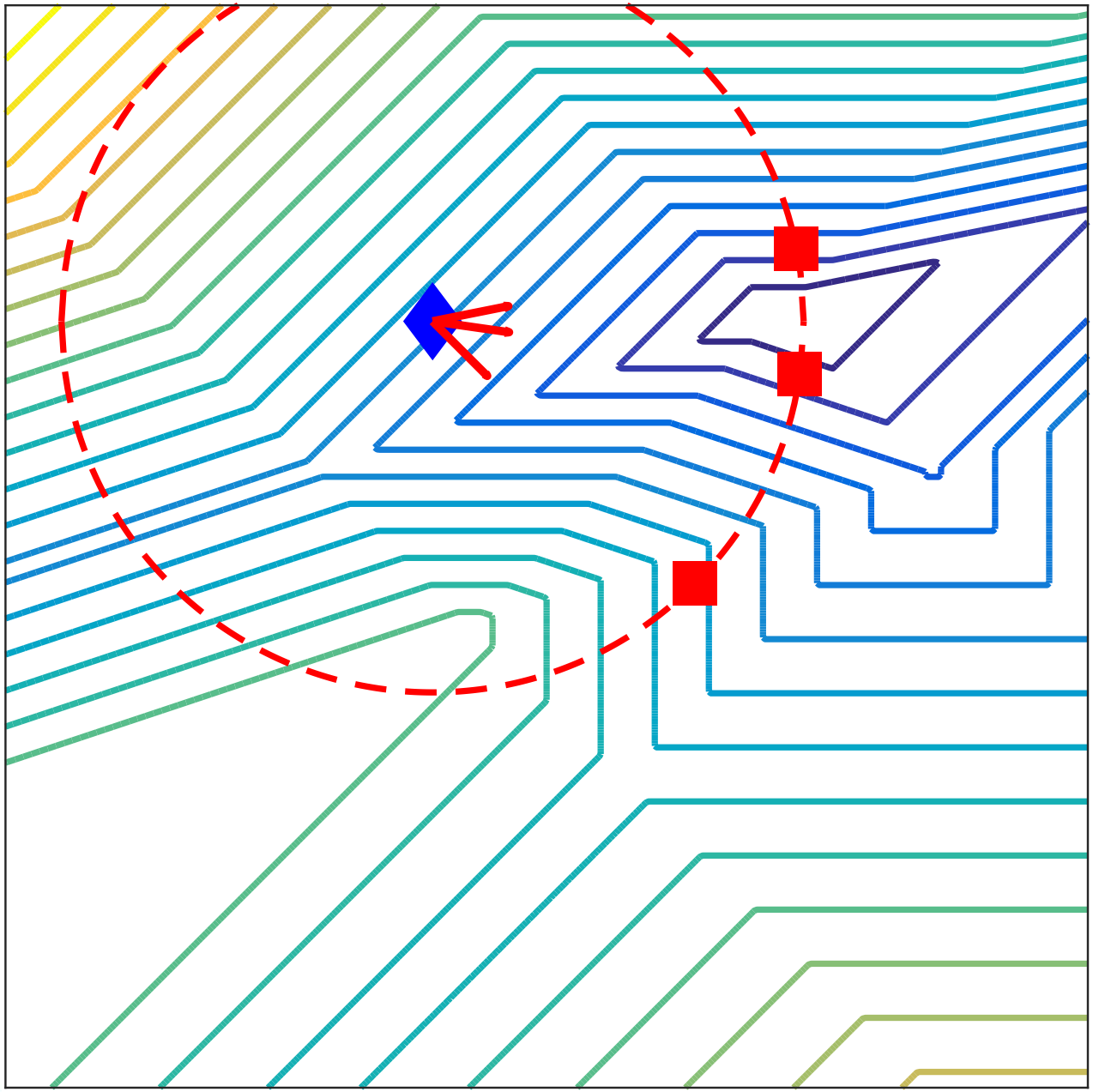
Manifold Sampling



Manifold Sampling



Manifold Sampling



Smooth master model

$$g^k \triangleq \mathbf{proj}\left(0, \mathbf{co}\left(\mathbb{G}^k\right)\right) \in \mathbf{co}\left(\mathbb{G}^k\right),$$

Smooth master model

$$g^k \triangleq \mathbf{proj}\left(0, \mathbf{co}\left(\mathbb{G}^k\right)\right) \in \mathbf{co}\left(\mathbb{G}^k\right),$$

where

$$\mathbb{G}^k \triangleq \bigcup_{i \in I_h(F(x^k))} \left\{\nabla M(x^k) \nabla h_i(F(x^k))\right\}$$

Smooth master model

$$g^k \triangleq \mathbf{proj}\left(0, \mathbf{co}\left(\mathbb{G}^k\right)\right) \in \mathbf{co}\left(\mathbb{G}^k\right),$$

where

$$\mathbb{G}^k \triangleq \bigcup_{i \in I_h(F(x^k))} \left\{\nabla M(x^k) \nabla h_i(F(x^k))\right\}$$

or

$$\mathbb{G}^k \triangleq \bigcup_{y \in Y} \bigcup_{i \in I_h(F(y))} \left\{\nabla M(x^k) \nabla h_i(F(x^k))\right\}$$

Smooth master model

$$g^k \triangleq \mathbf{proj}\left(0, \mathbf{co}\left(\mathbb{G}^k\right)\right) \in \mathbf{co}\left(\mathbb{G}^k\right),$$

where

$$\mathbb{G}^k \triangleq \bigcup_{i \in I_h(F(x^k))} \left\{\nabla M(x^k) \nabla h_i(F(x^k))\right\}$$

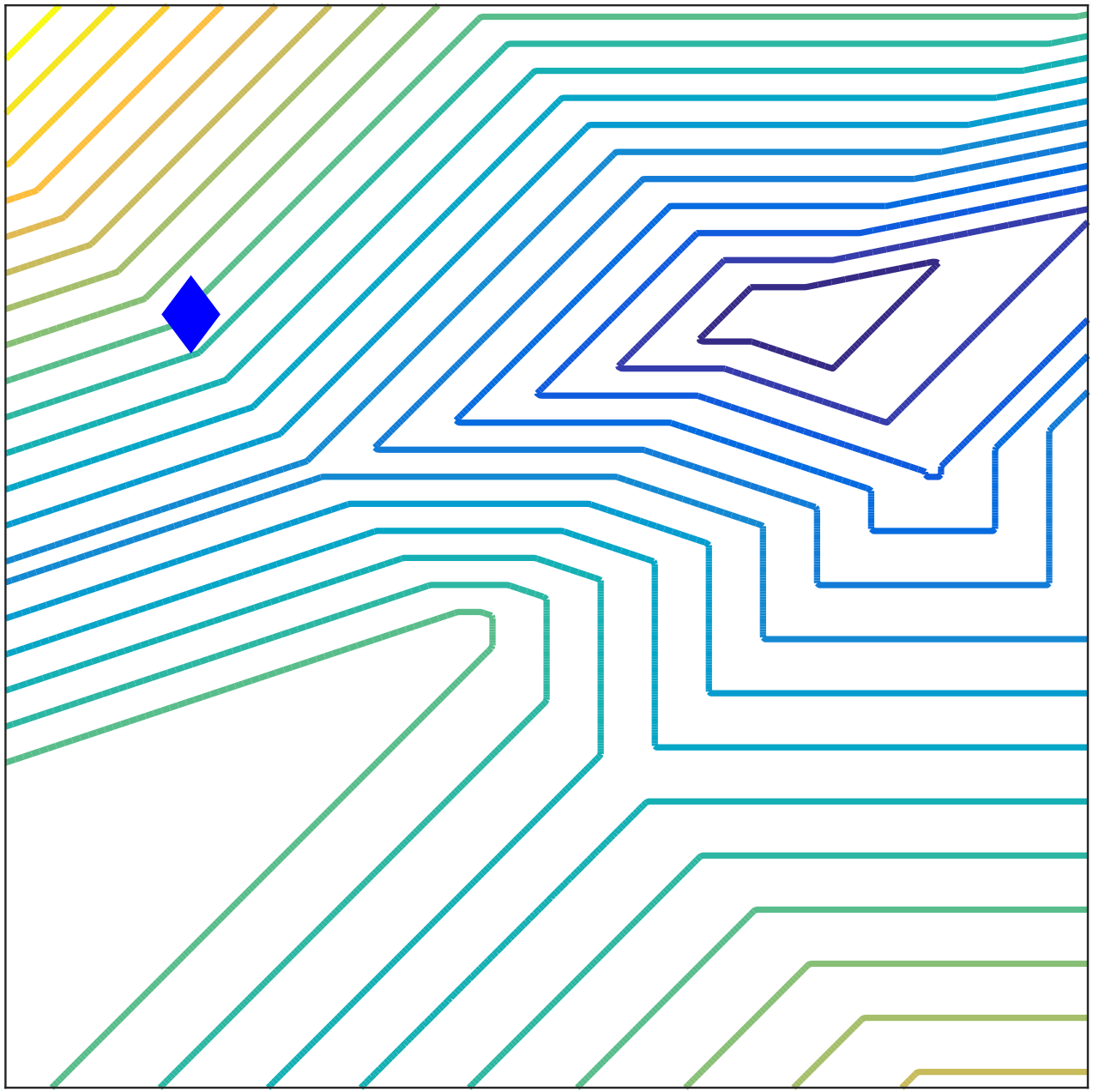
or

$$\mathbb{G}^k \triangleq \bigcup_{y \in Y} \bigcup_{i \in I_h(F(y))} \left\{\nabla M(x^k) \nabla h_i(F(x^k))\right\}$$

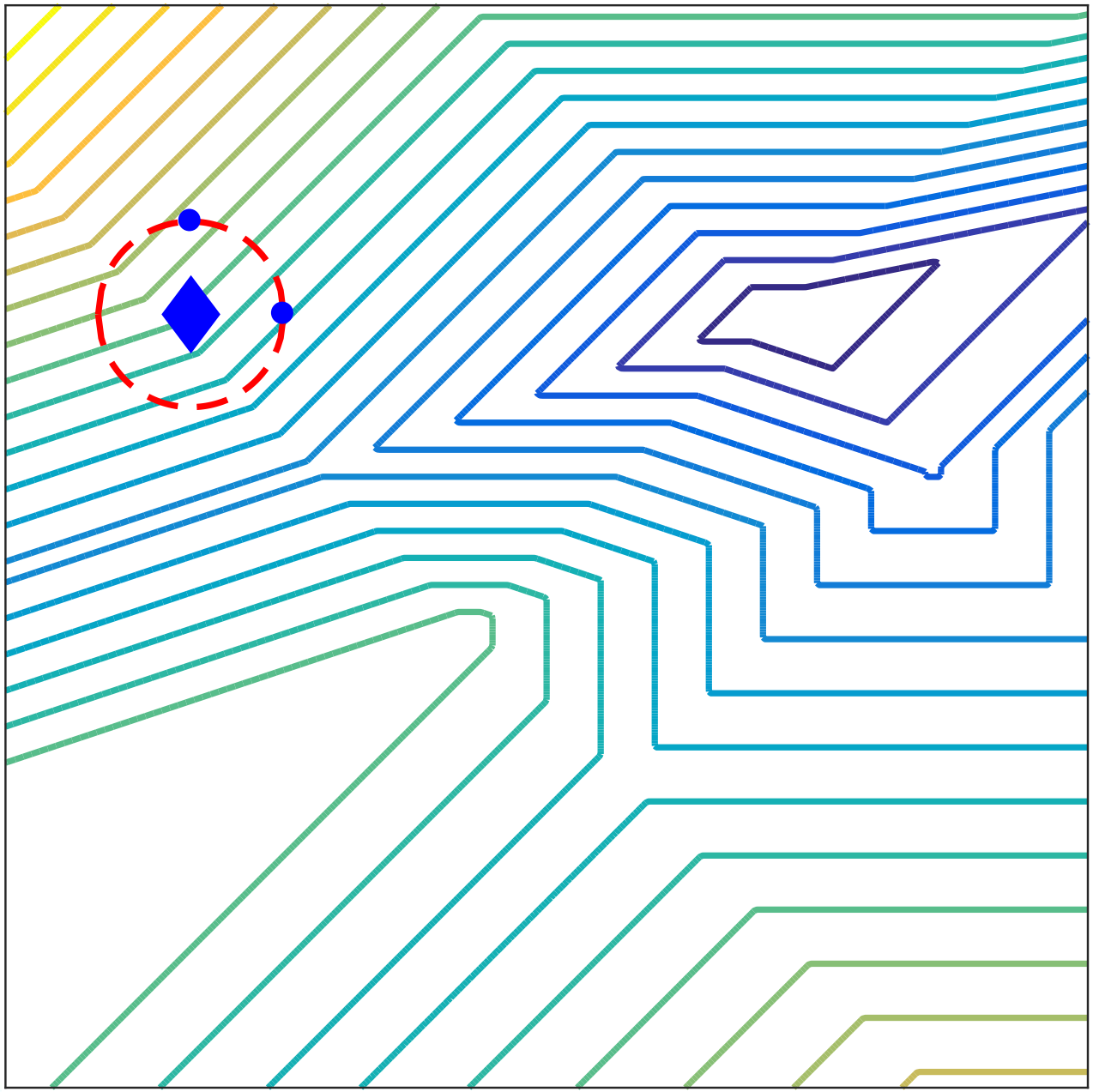
Define the smooth *master model* $m_k^f : \mathbb{R}^n \to \mathbb{R}$ (with gradient g^k) and obtain step by (approximately) solving

$$\begin{aligned} &\underset{s}{\text{minimize}}\ m_k^f(x^k + s) \\ &\text{subject to: } s \in \mathcal{B}(0, \Delta_k) \end{aligned}$$

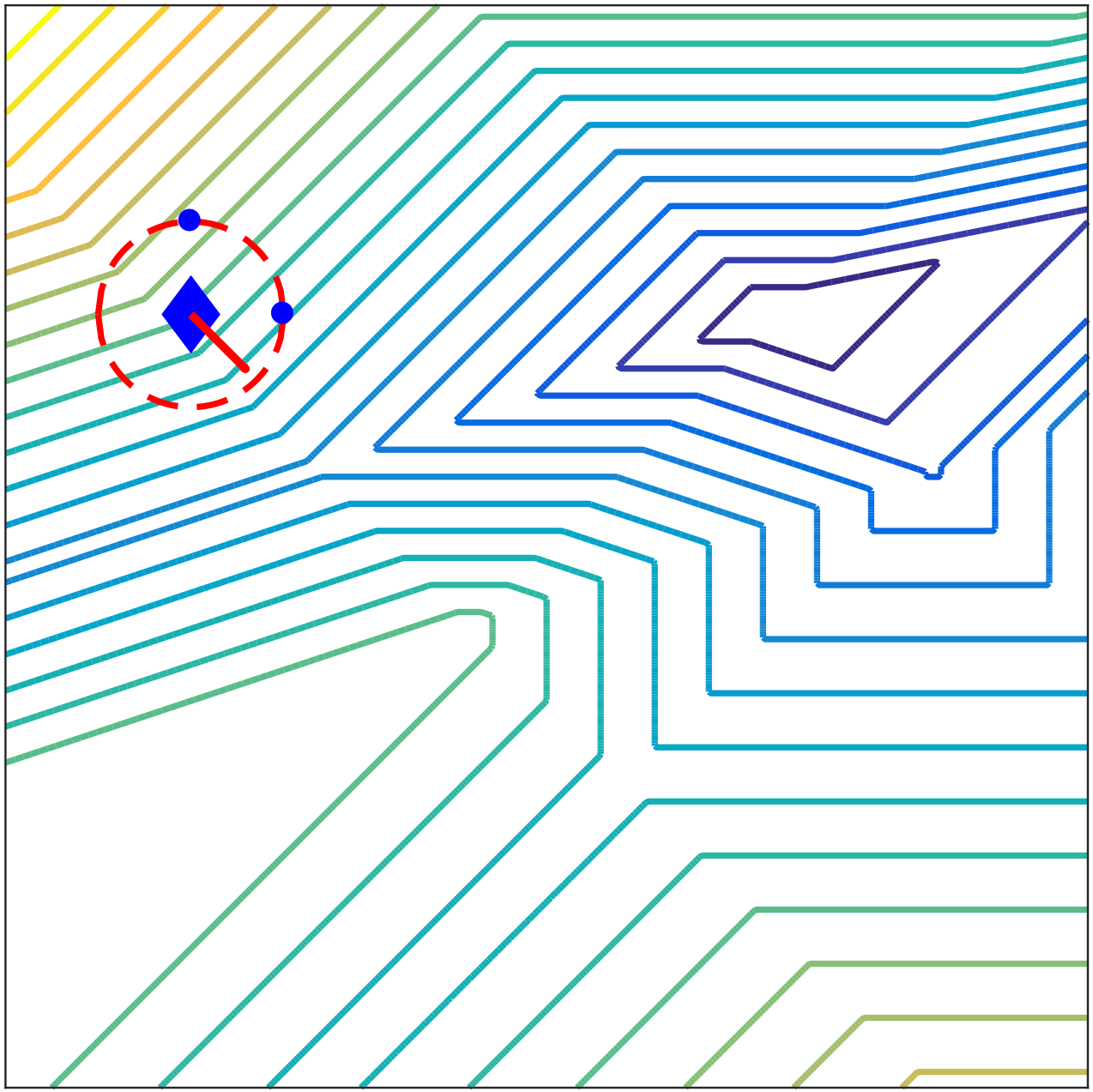
Manifold Sampling



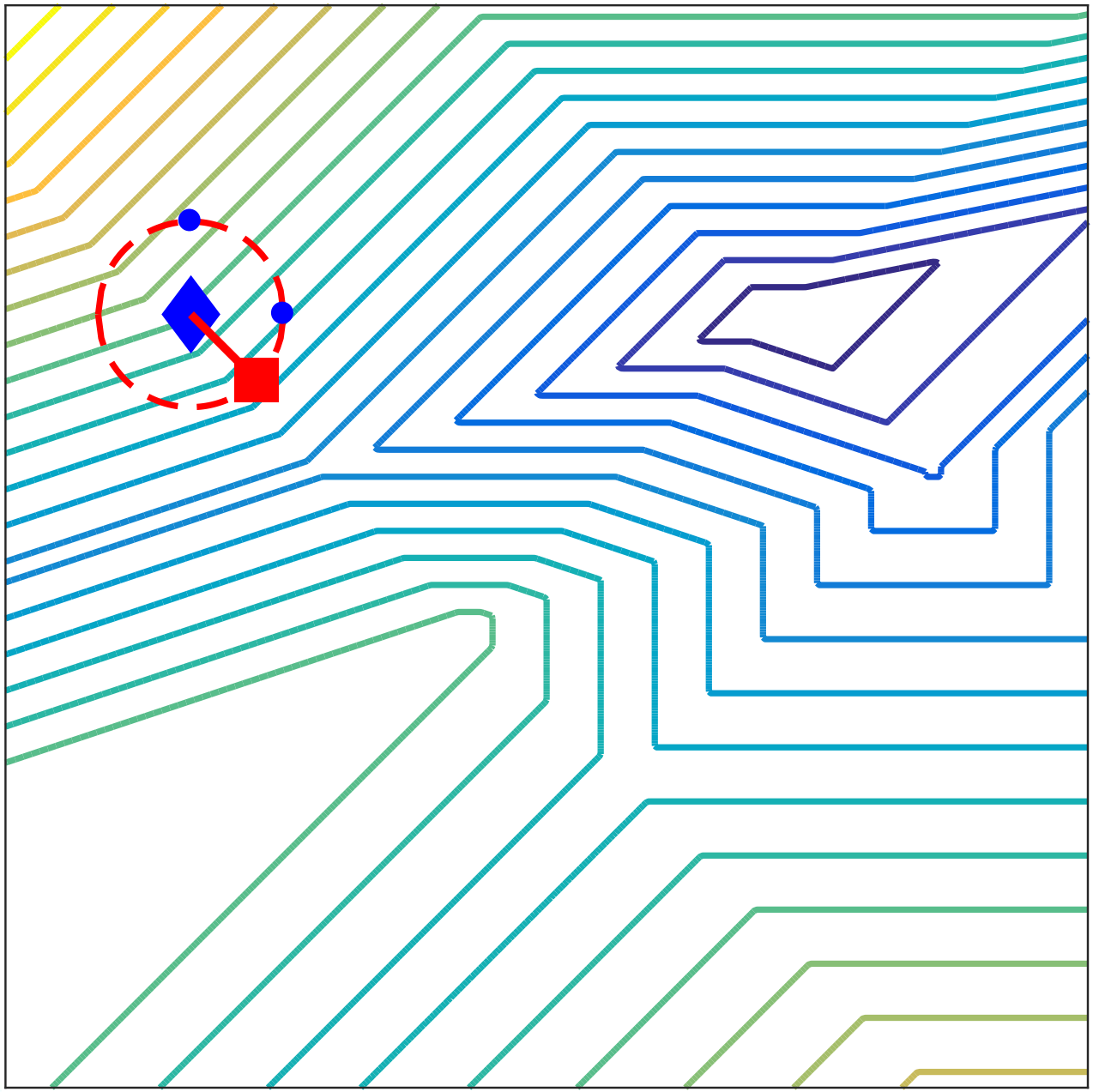
Manifold Sampling



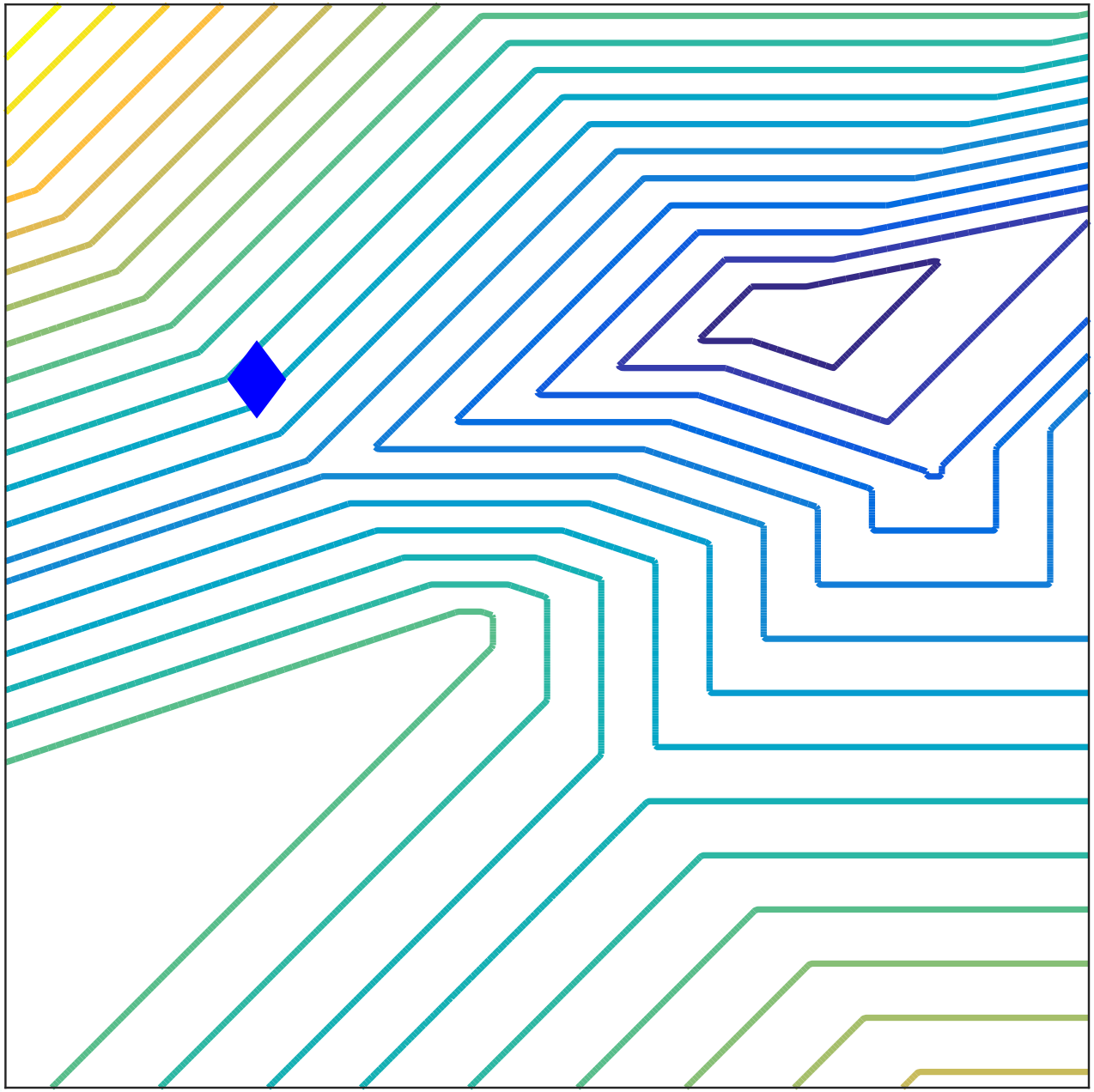
Manifold Sampling



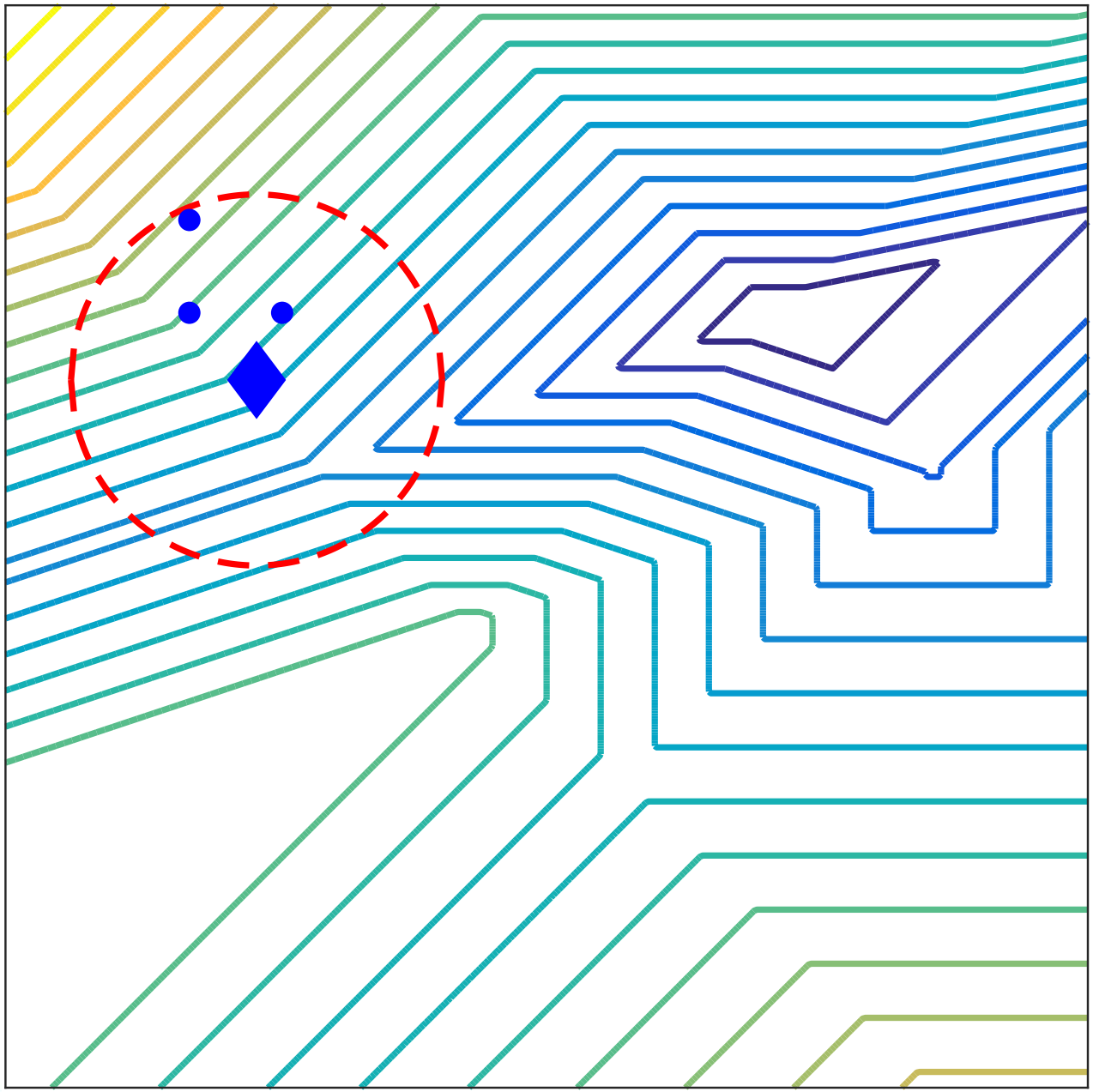
Manifold Sampling



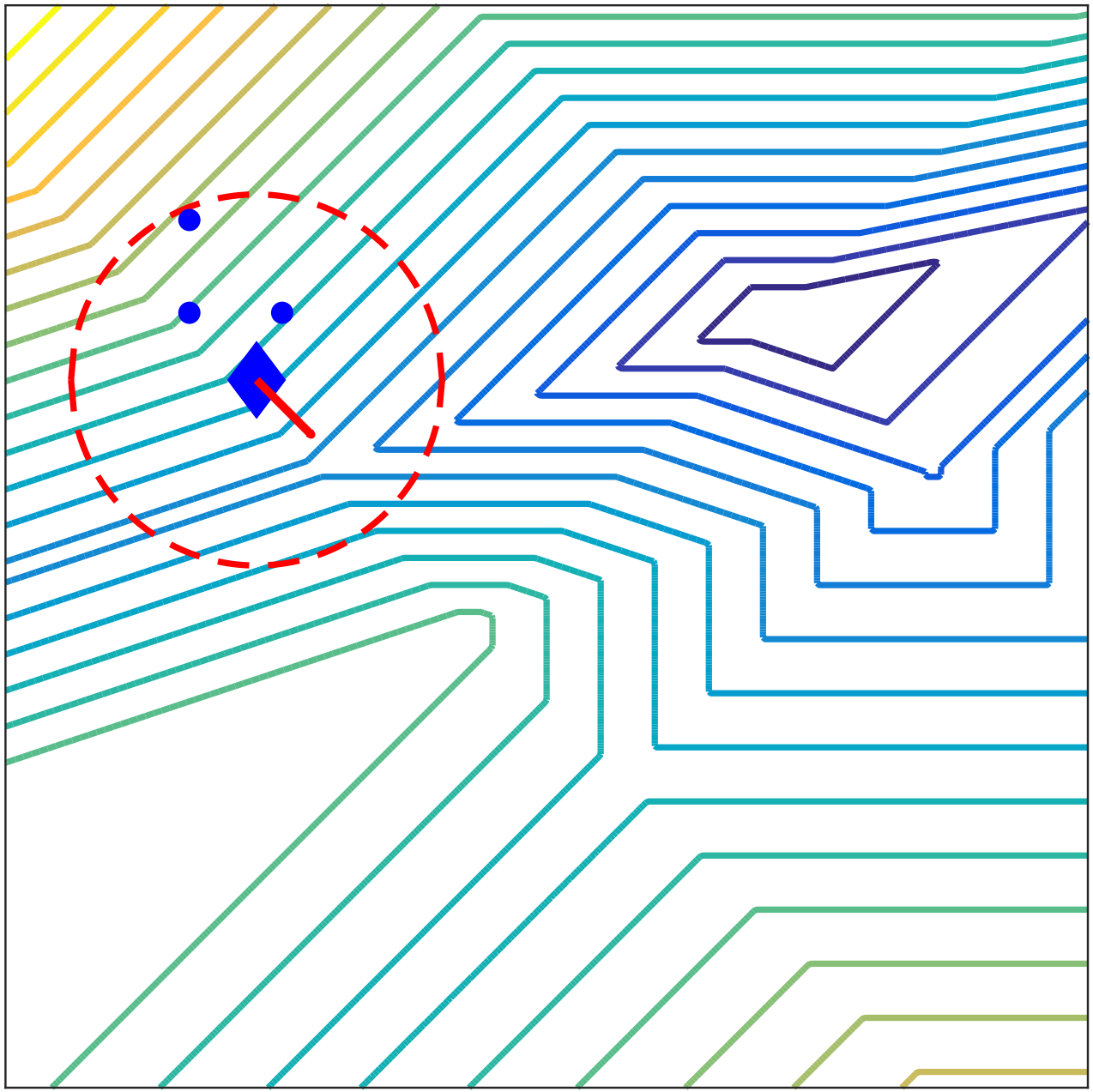
Manifold Sampling



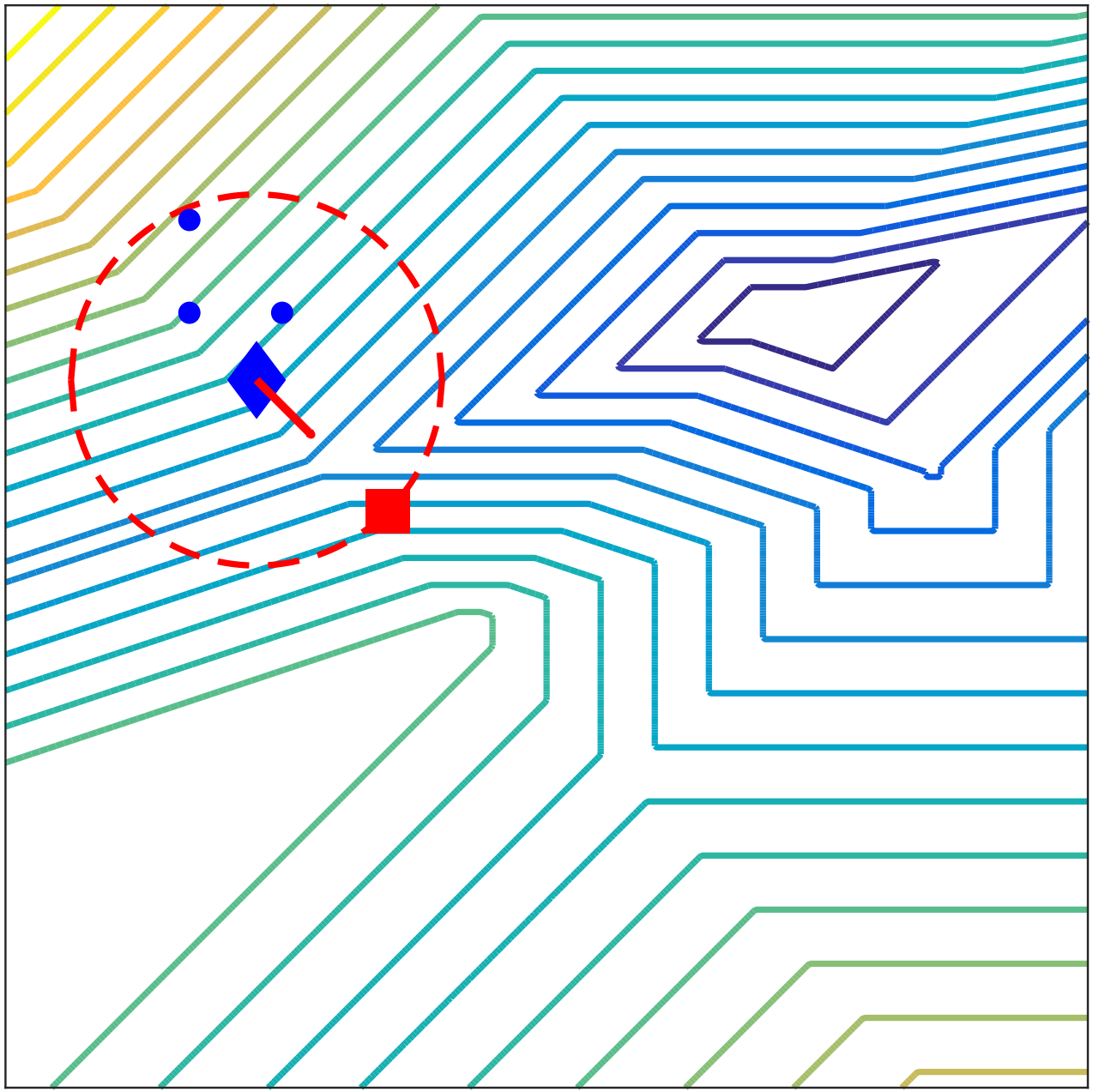
Manifold Sampling



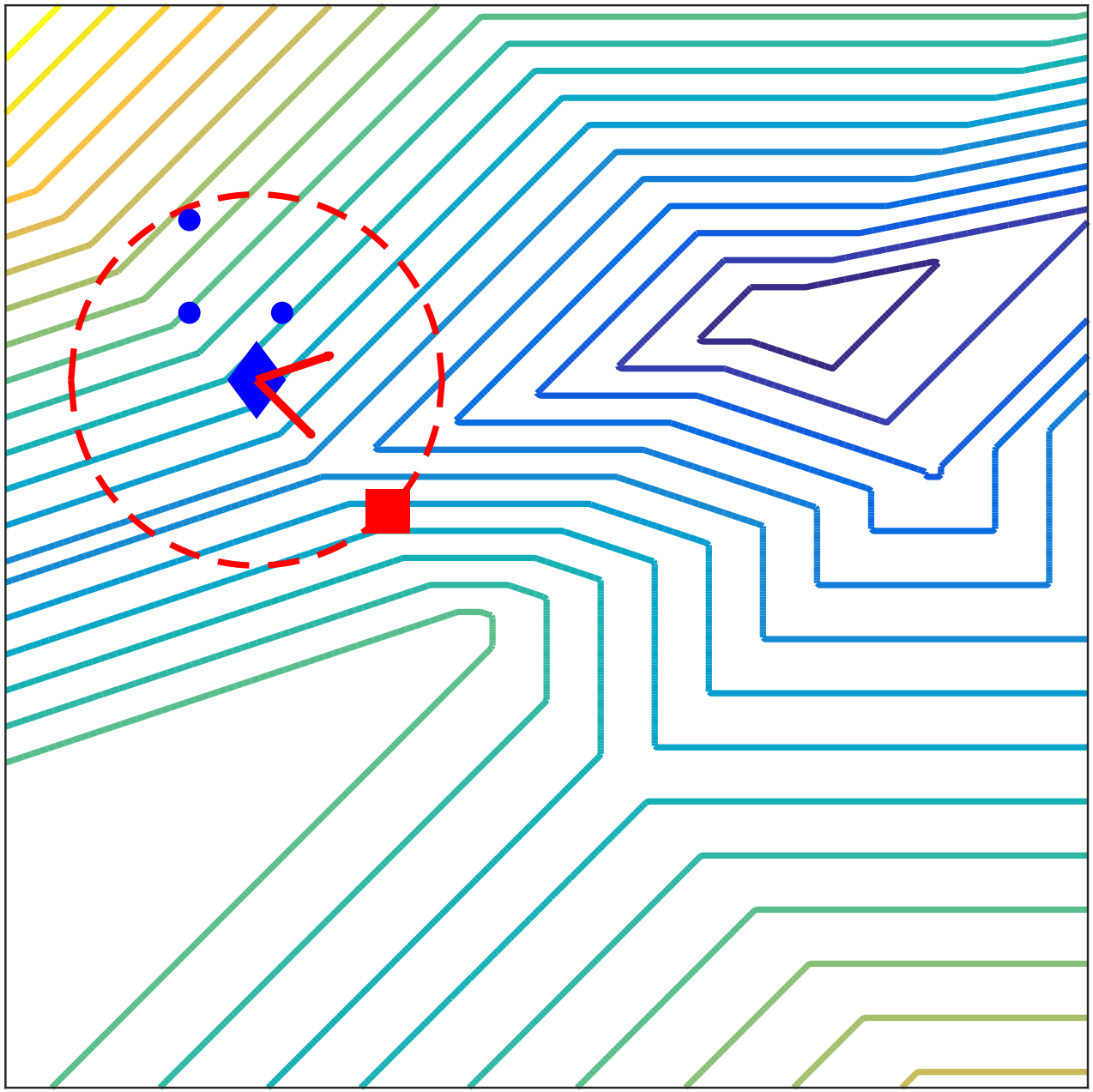
Manifold Sampling



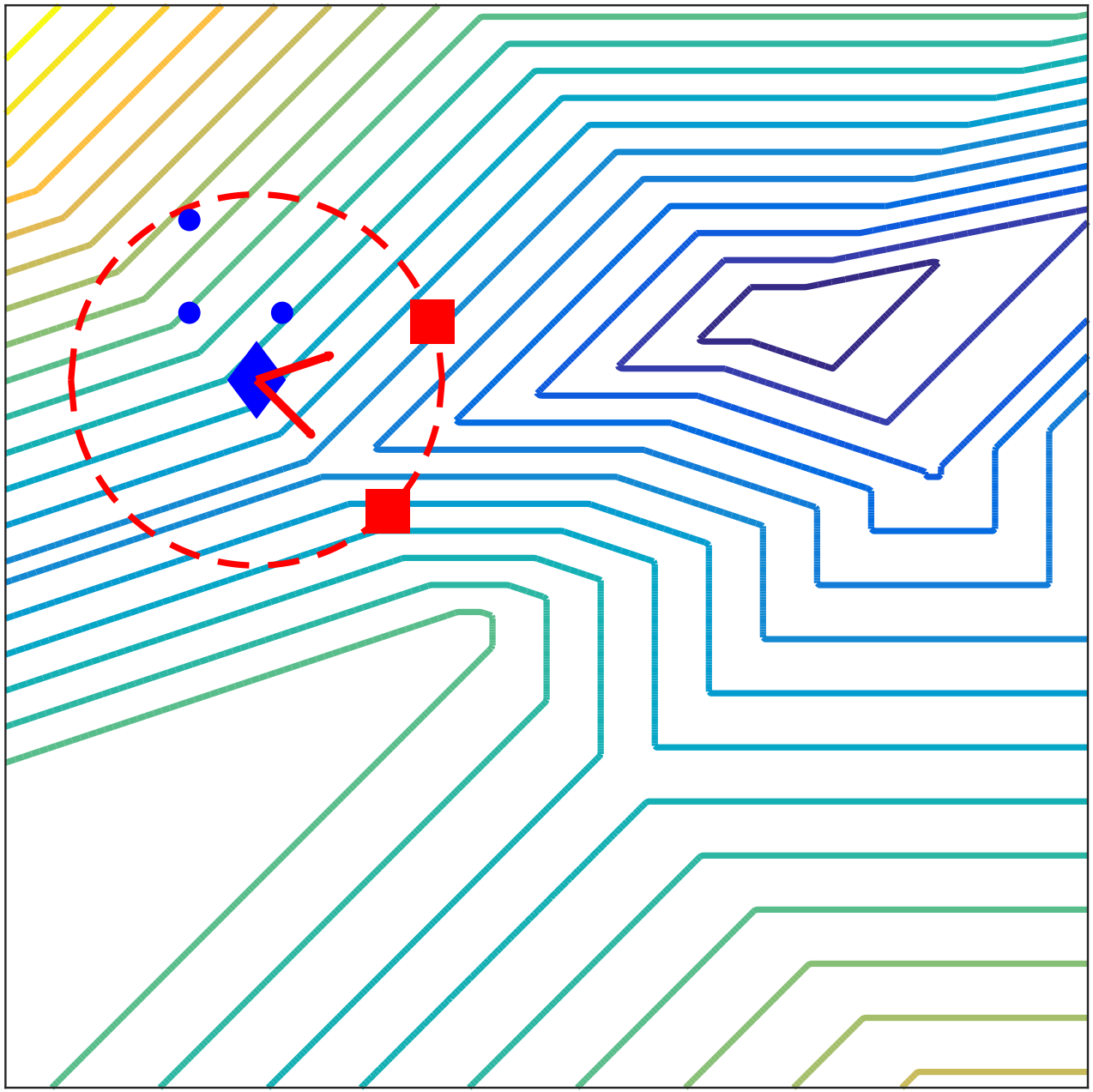
Manifold Sampling



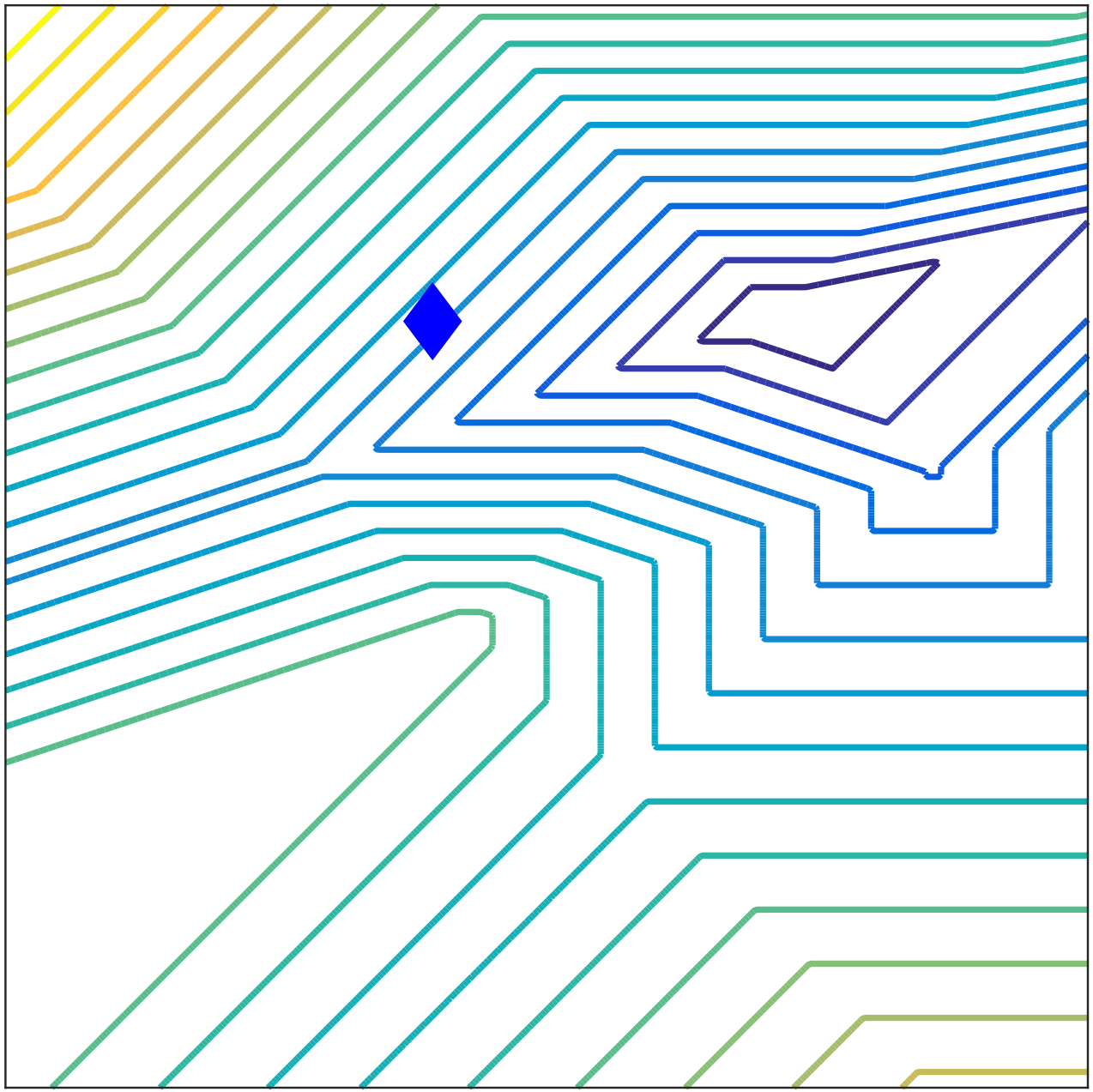
Manifold Sampling



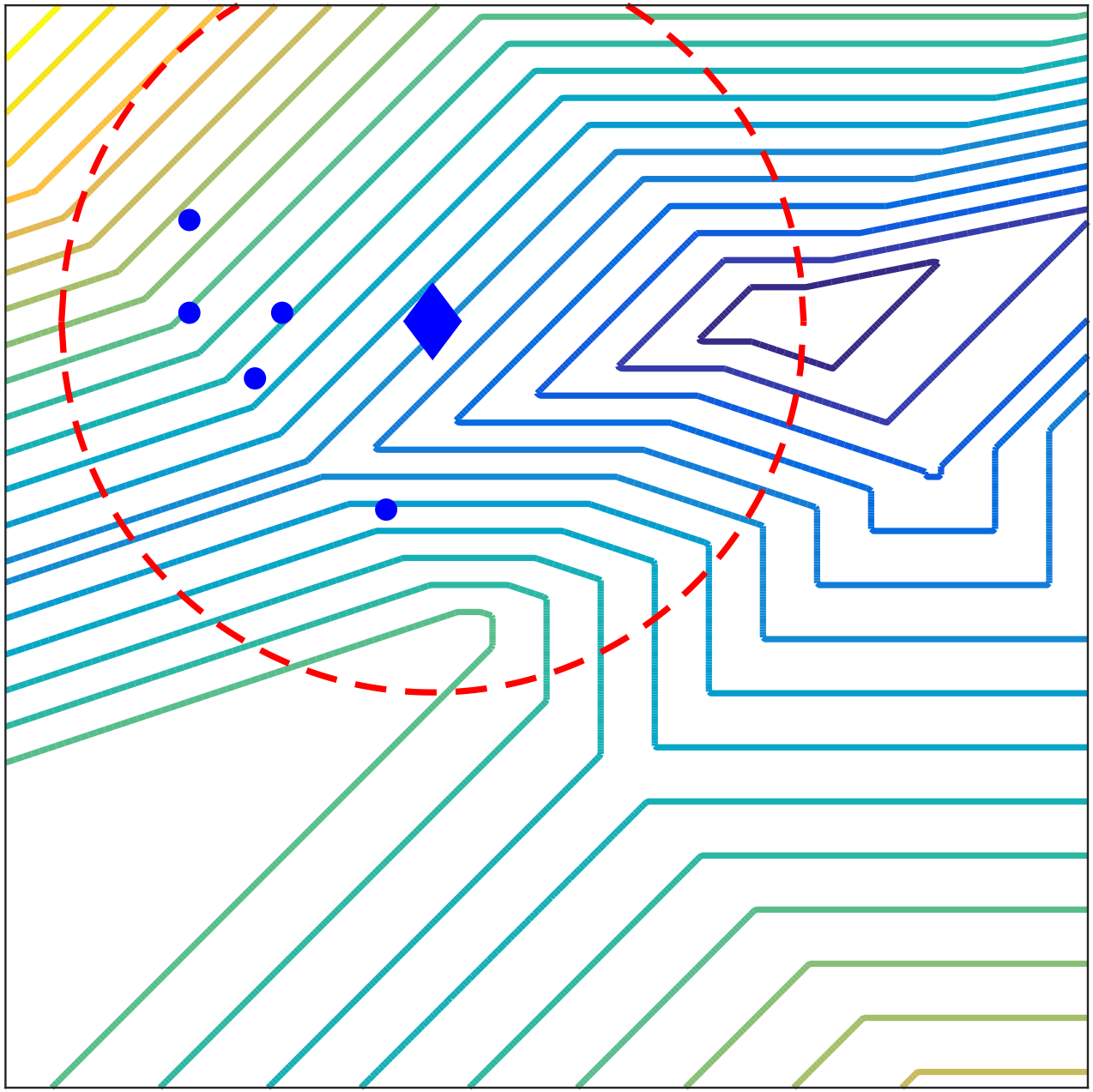
Manifold Sampling



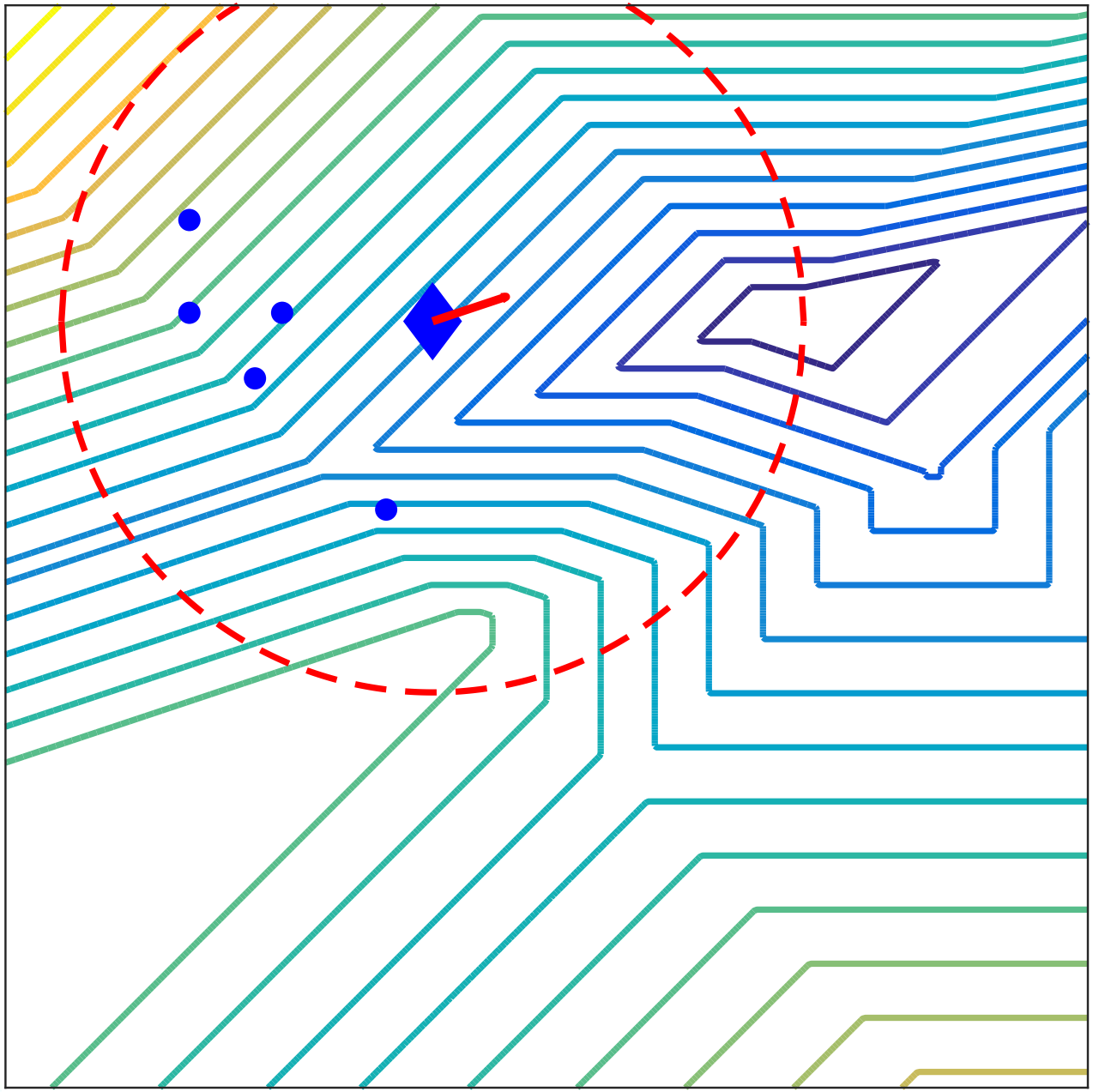
Manifold Sampling



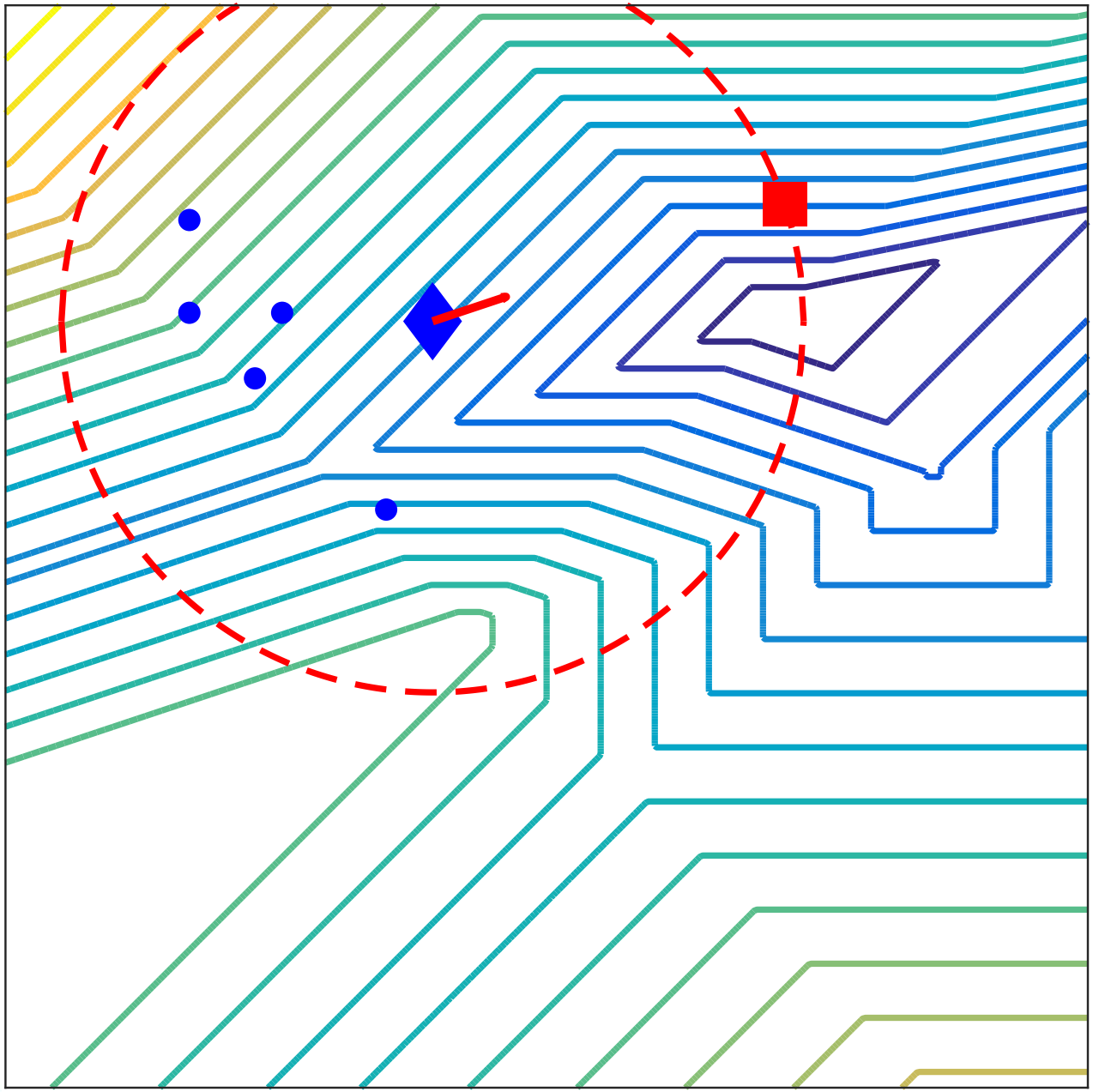
Manifold Sampling



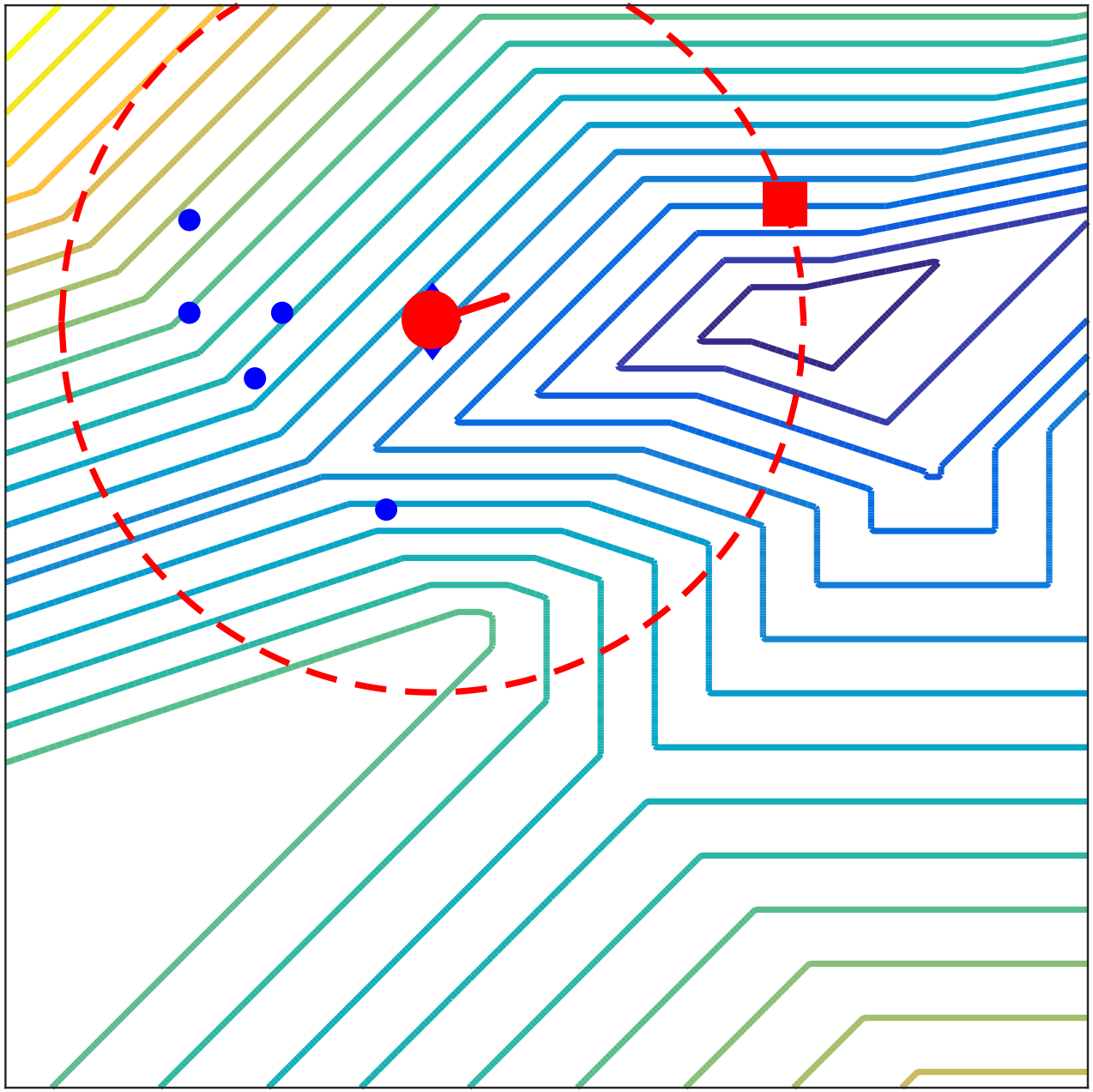
Manifold Sampling



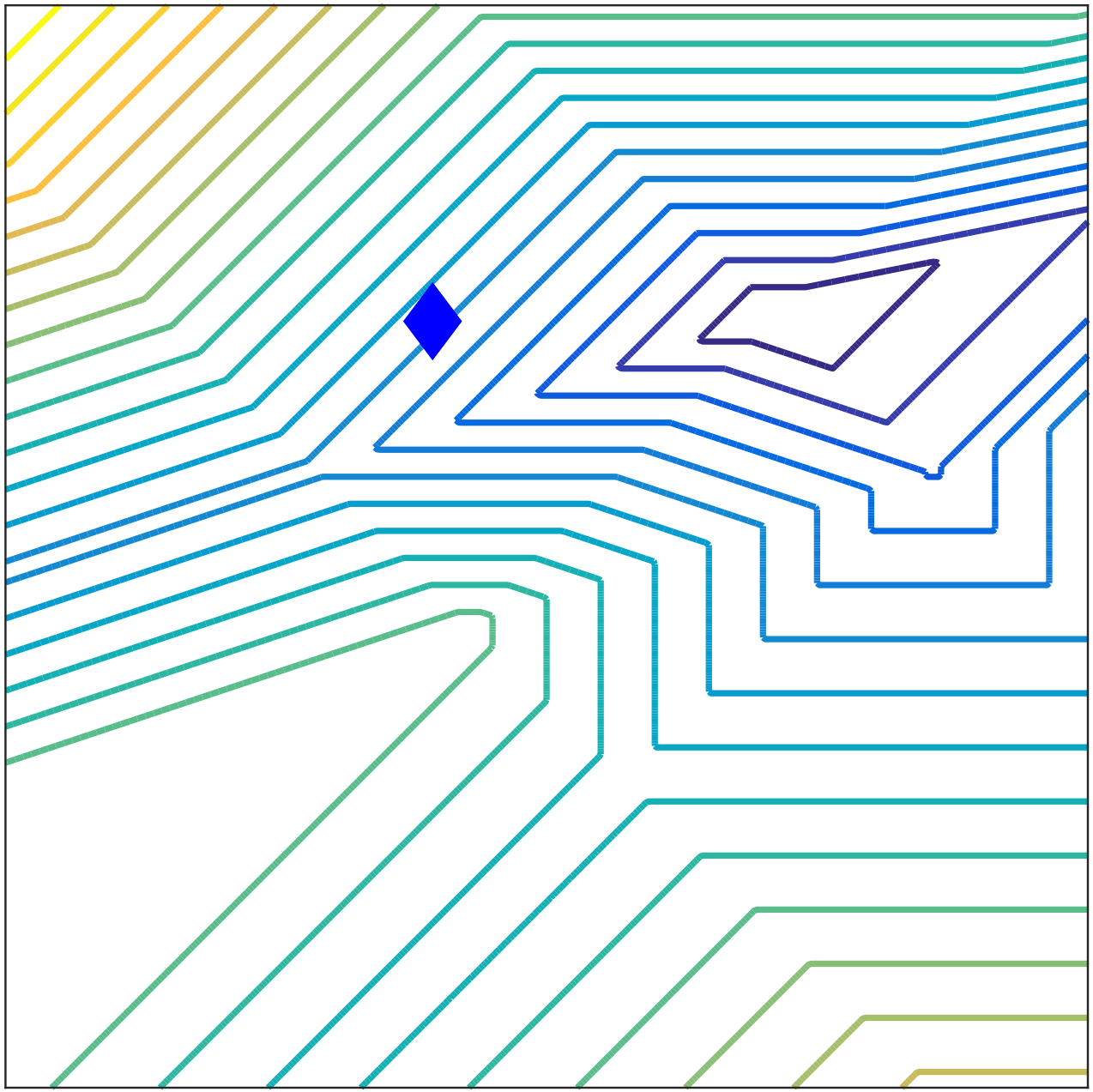
Manifold Sampling



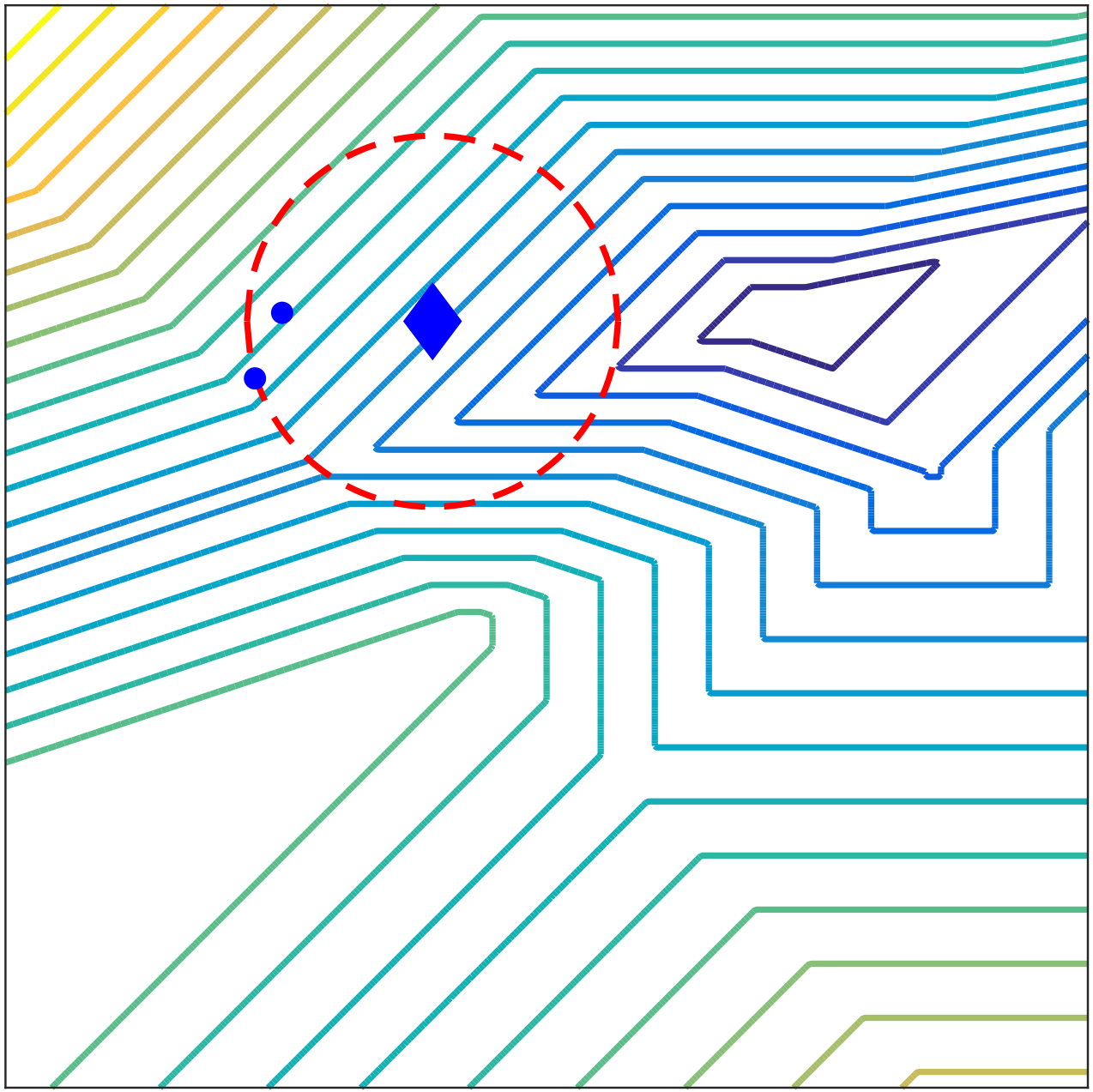
Manifold Sampling



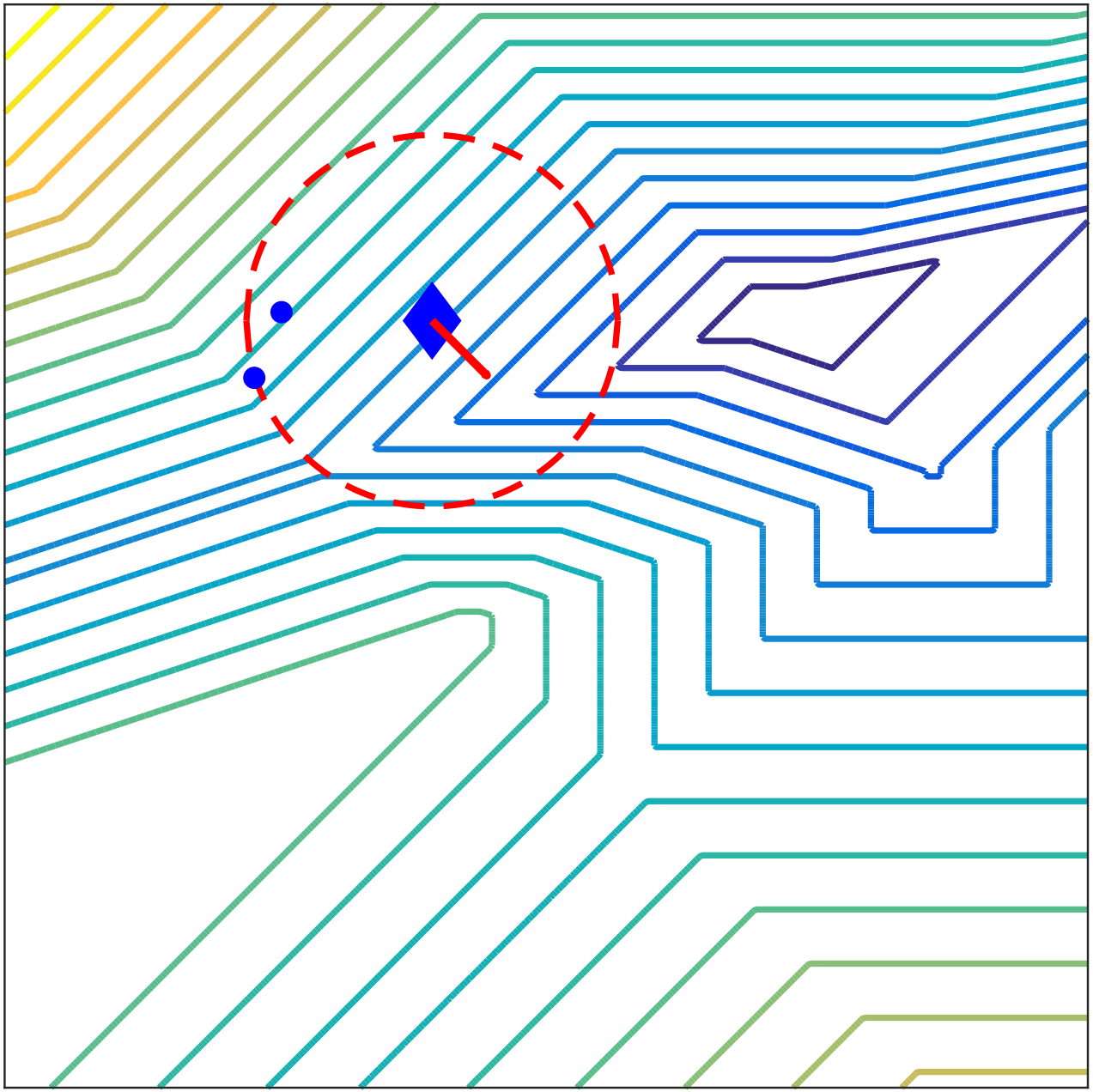
Manifold Sampling



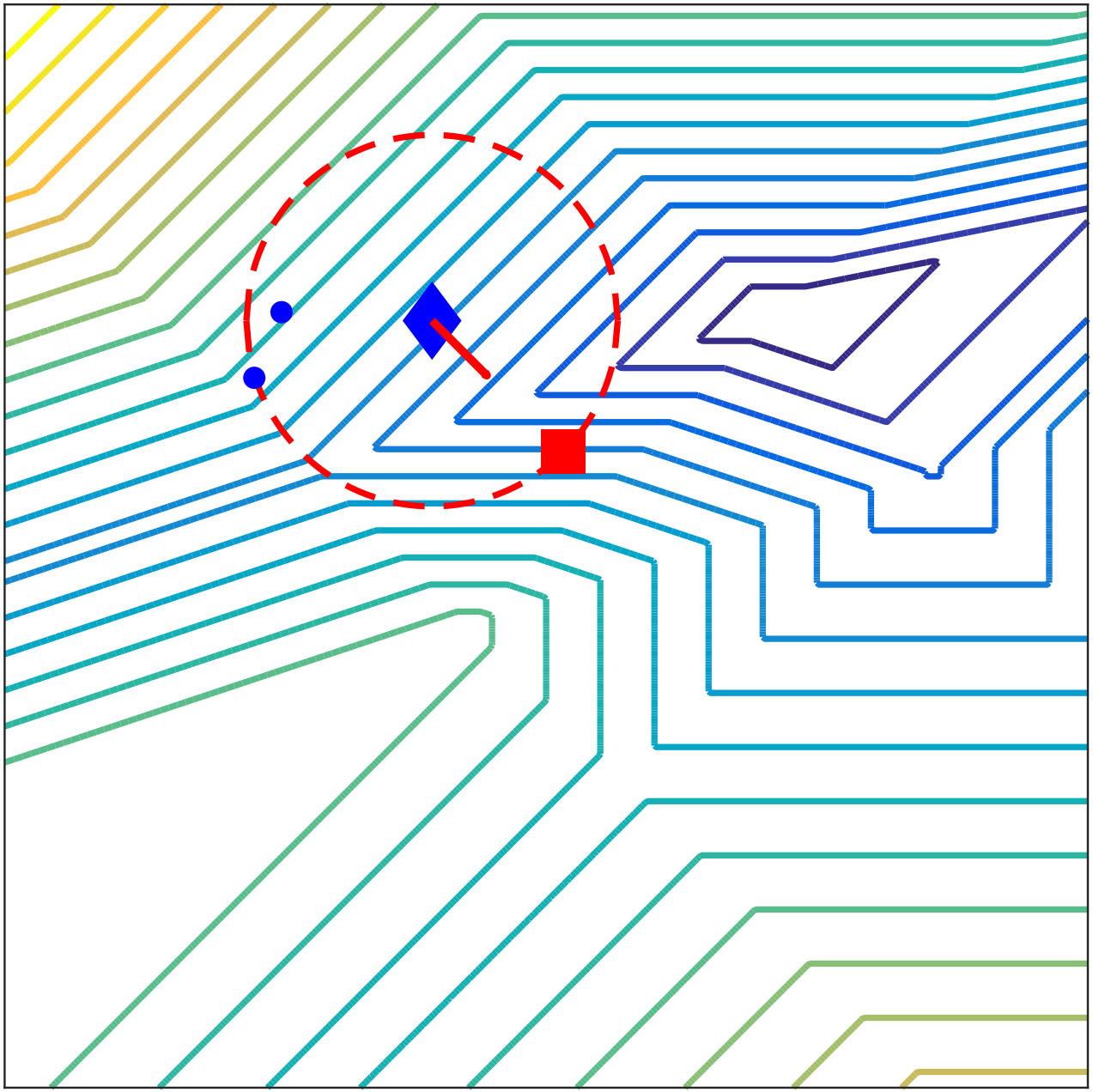
Manifold Sampling



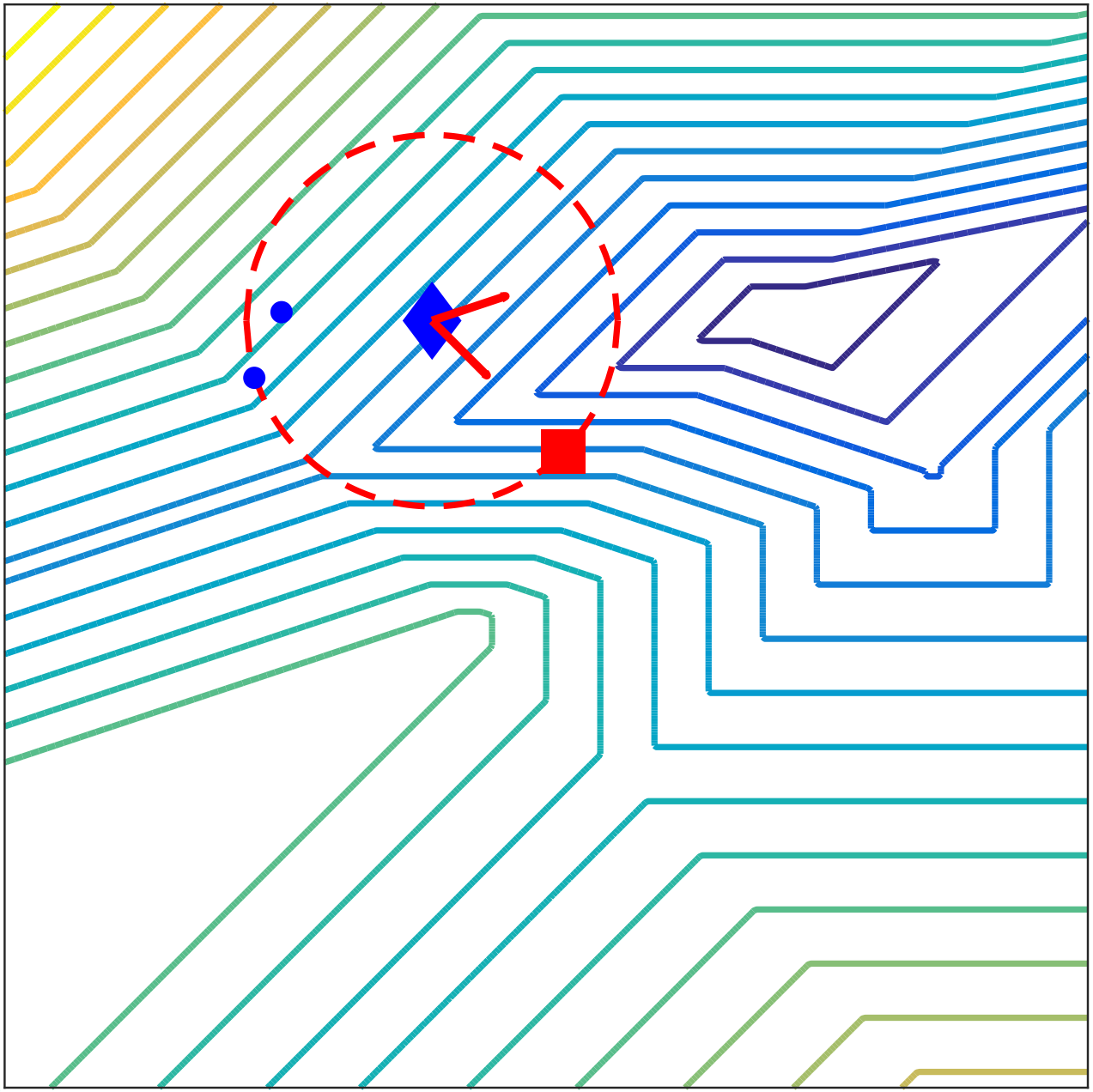
Manifold Sampling



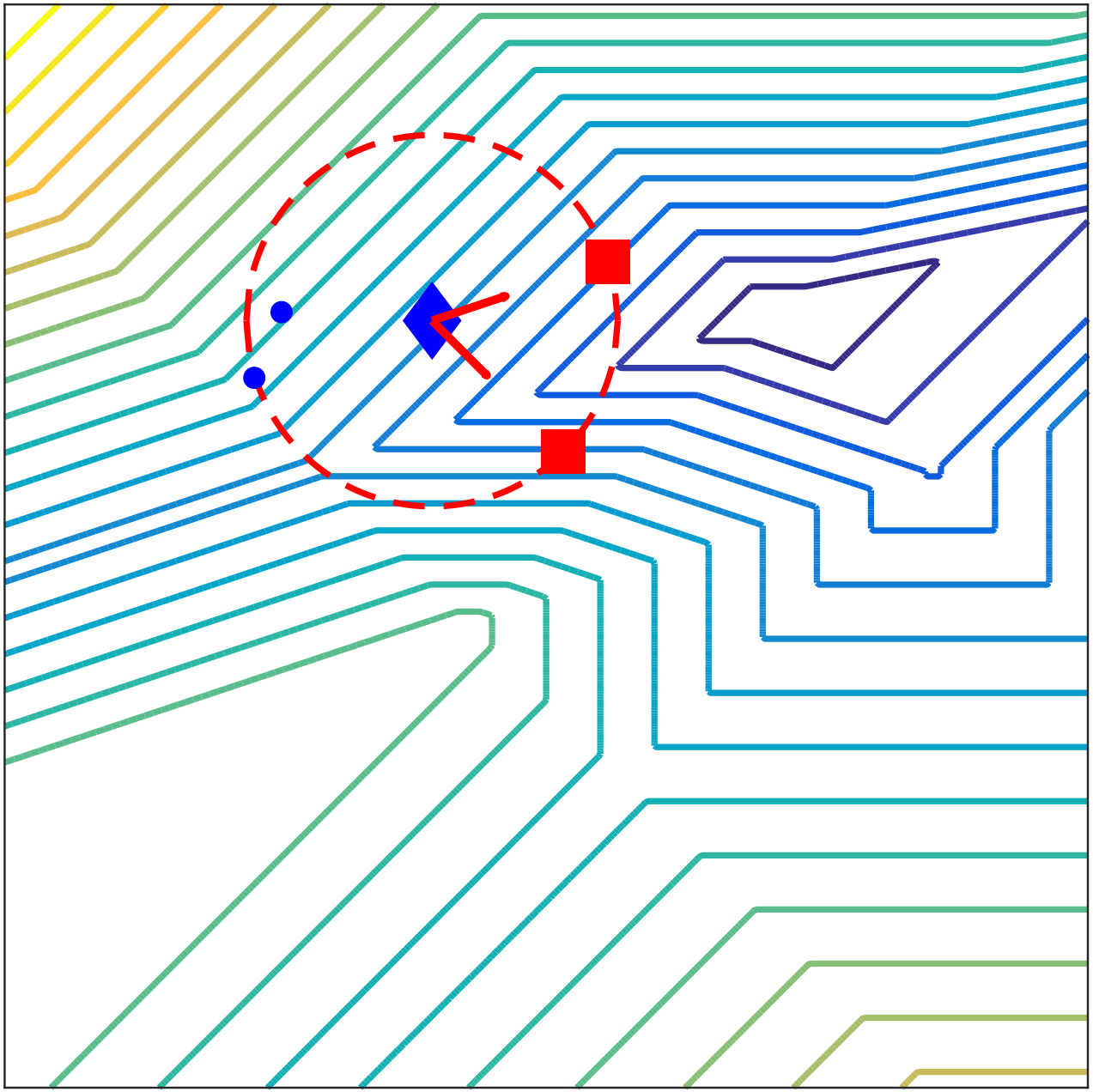
Manifold Sampling



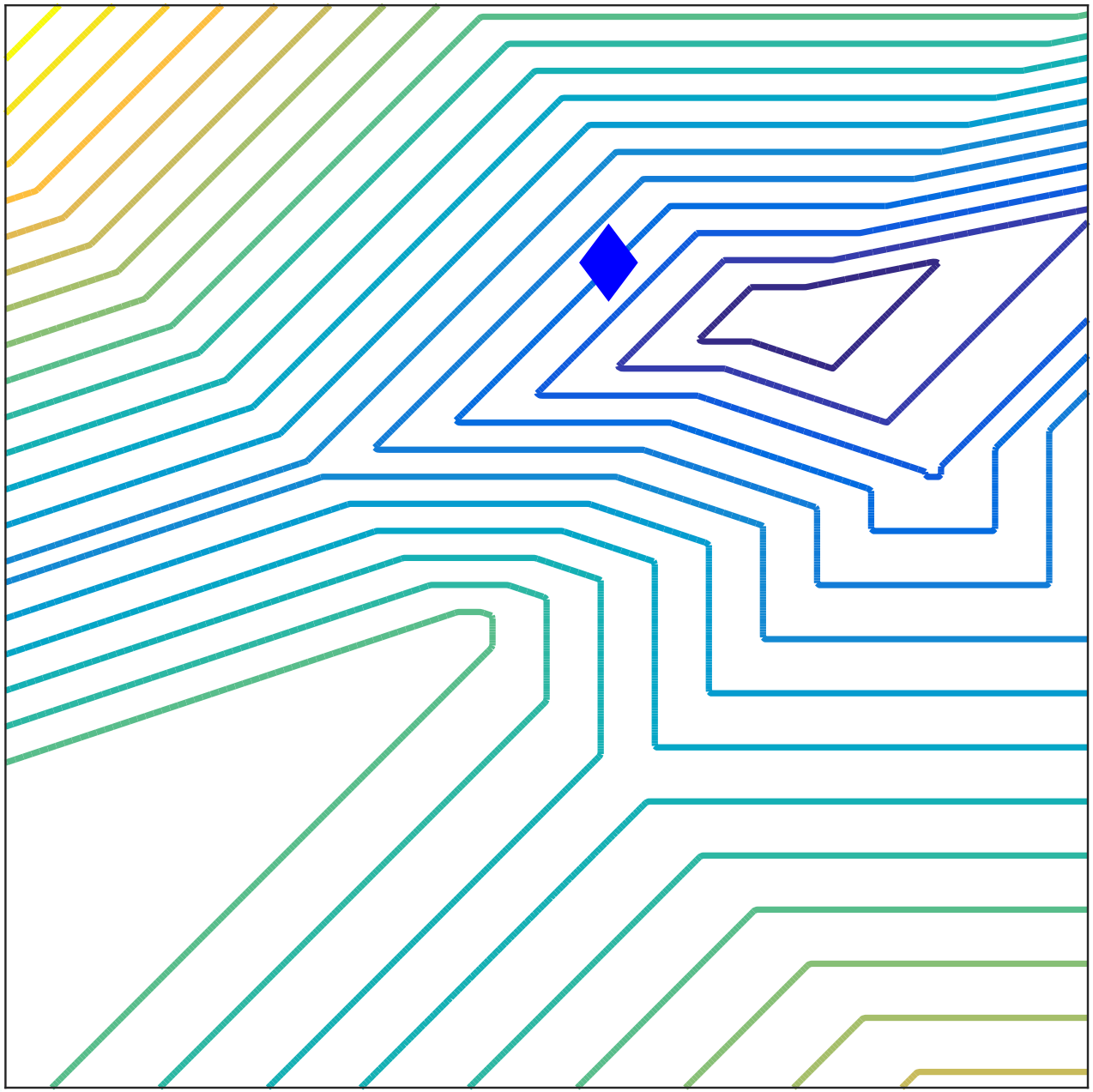
Manifold Sampling



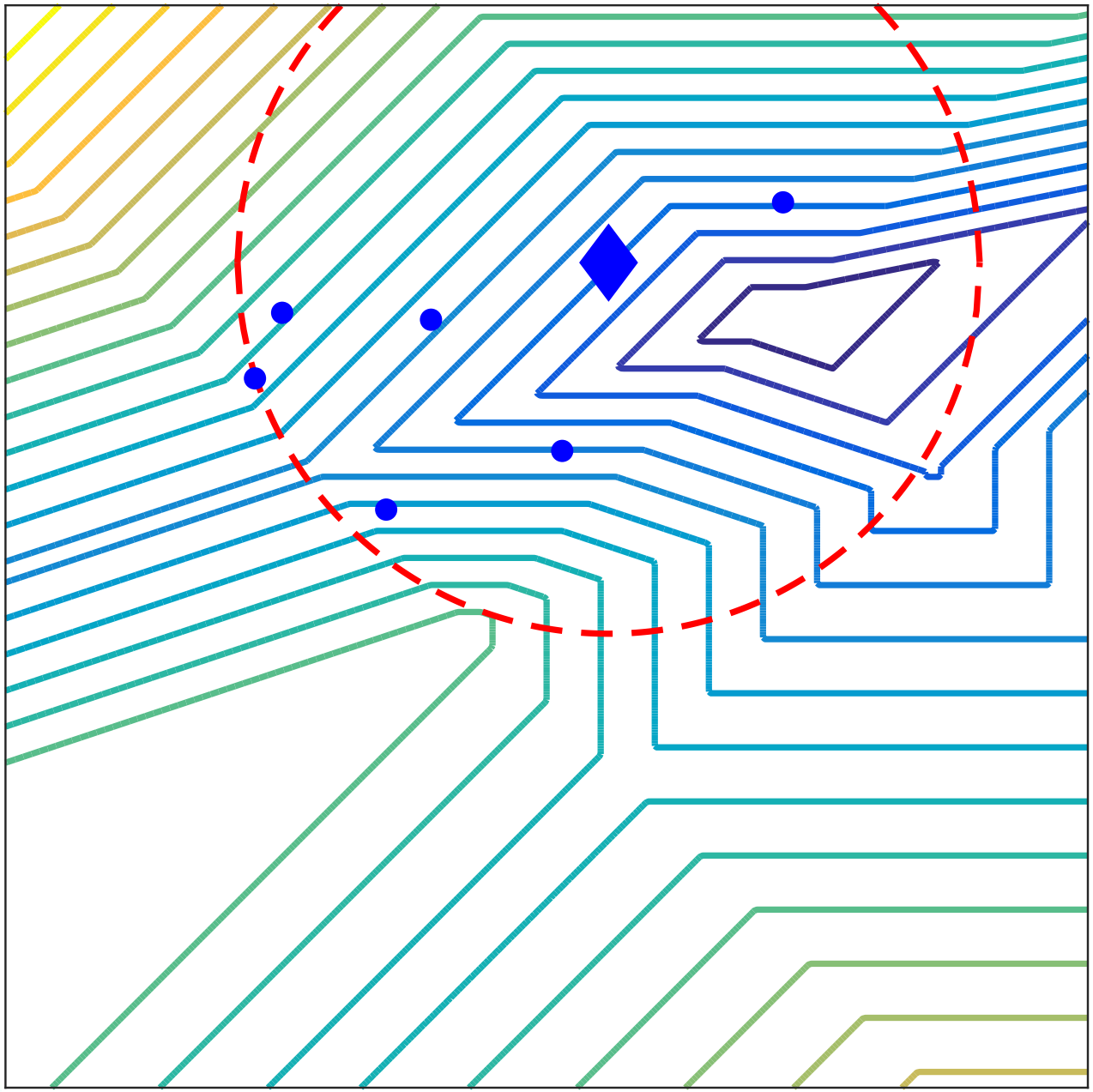
Manifold Sampling



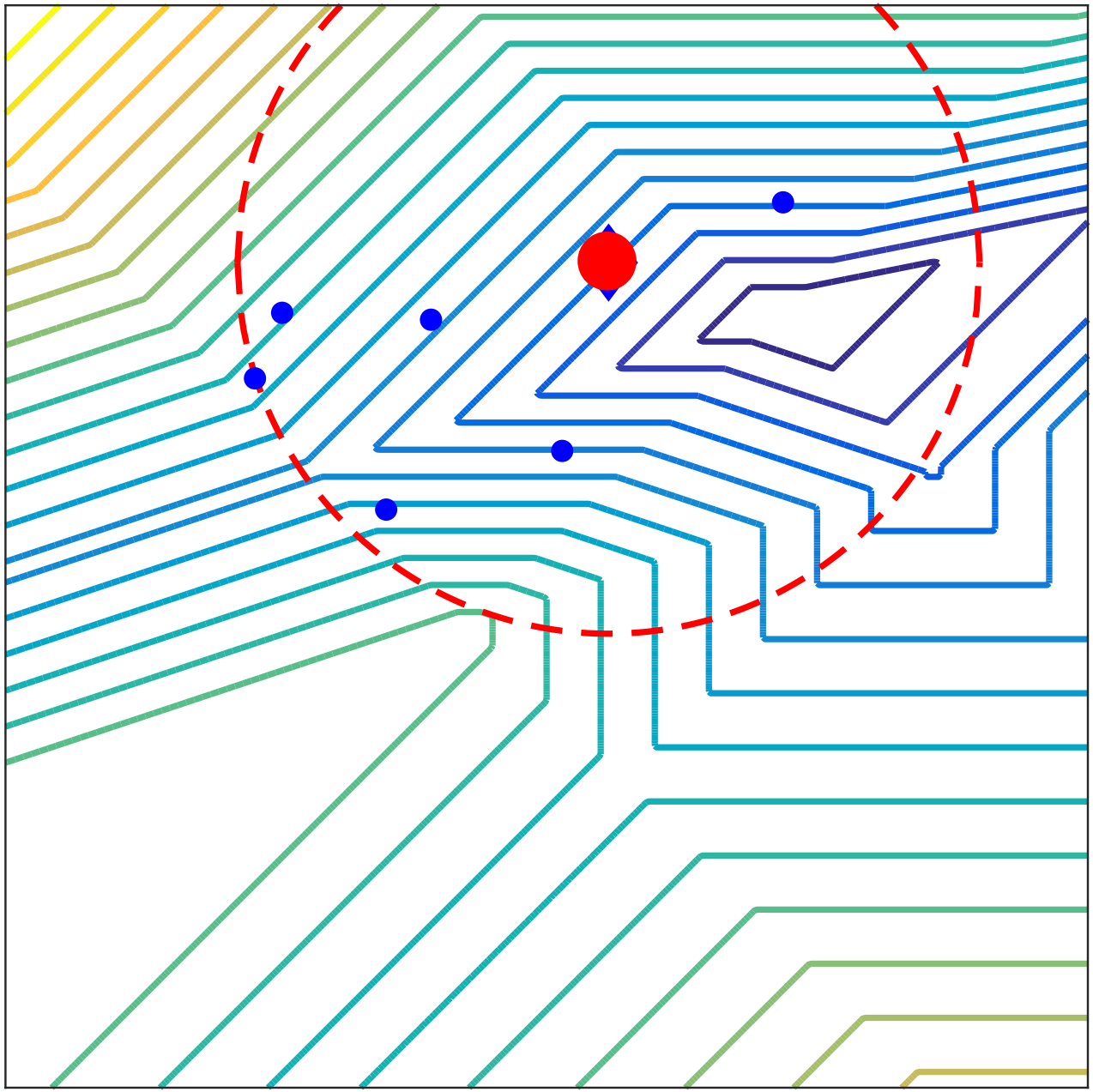
Manifold Sampling



Manifold Sampling



Manifold Sampling



Better trust-region subproblem?

Instead of solving

$$\underset{s}{\text{minimize}}\ m_k^f(x^k + s)$$

$$\text{subject to: } s \in \mathcal{B}(0, \Delta_k)$$

How about

$$\underset{s}{\text{minimize}}\ h(M(x^k + s))$$

$$\text{subject to: } s \in \mathcal{B}(0, \Delta_k)$$

Better trust-region subproblem?

Instead of solving

$$\begin{aligned} &\underset{s}{\text{minimize}} \; m_k^f(x^k + s) \\ &\text{subject to: } s \in \mathcal{B}(0, \Delta_k) \end{aligned}$$

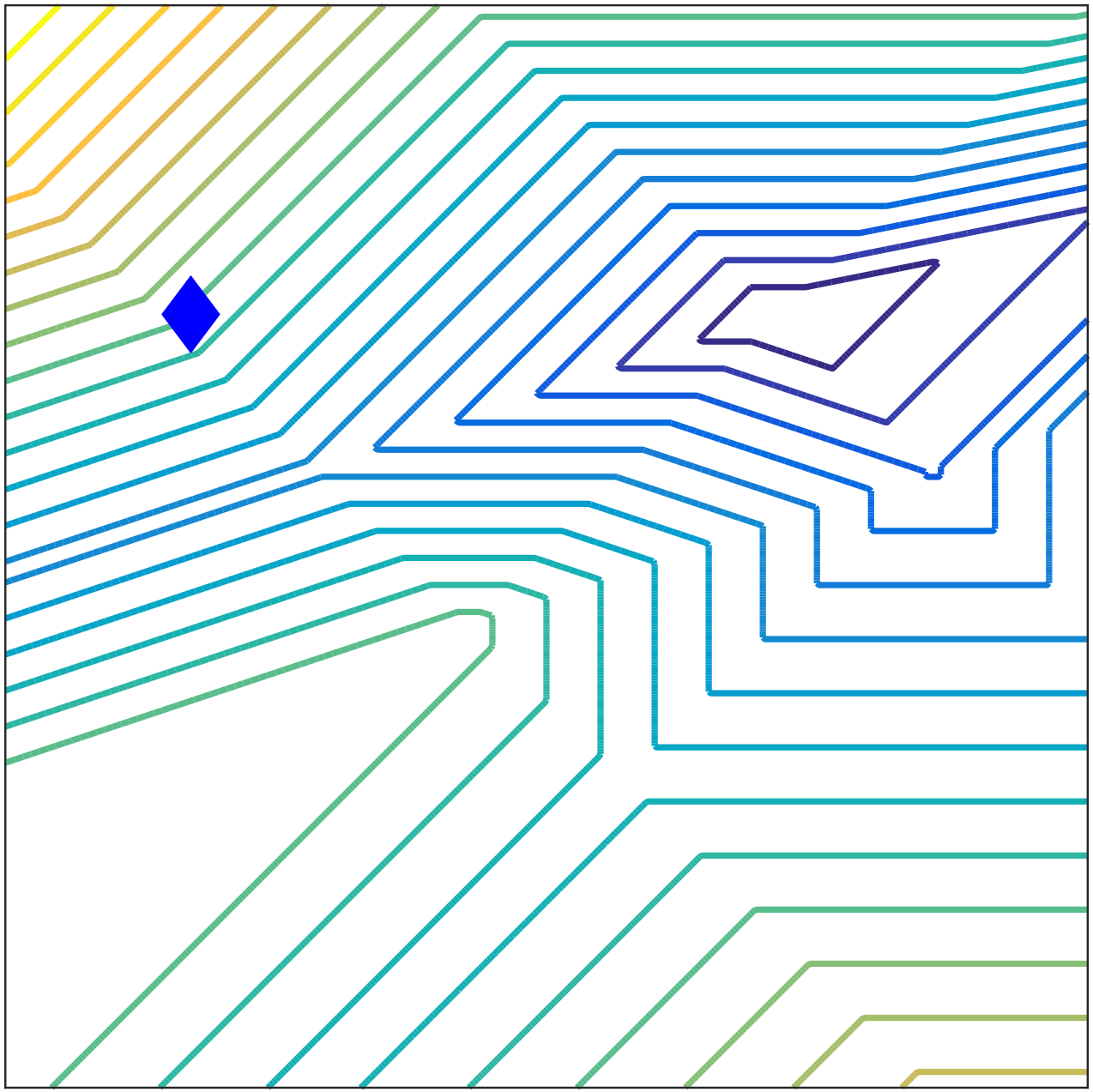
How about

$$\begin{aligned} &\underset{s}{\text{minimize}} \; h(M(x^k + s)) \\ &\text{subject to: } s \in \mathcal{B}(0, \Delta_k) \end{aligned}$$

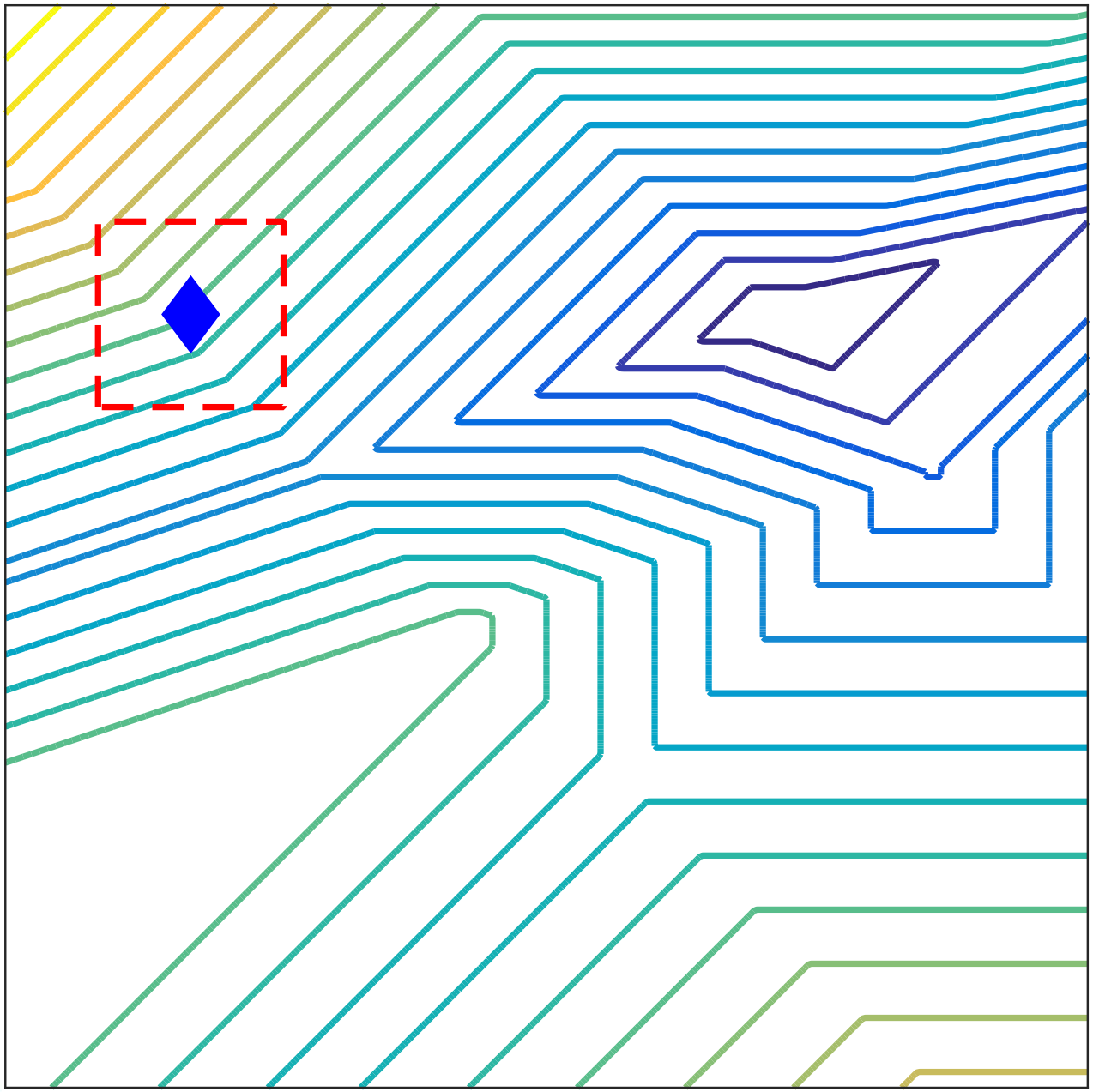
For censored ℓ_1 loss:

$$\begin{aligned} &\underset{s}{\text{minimize}} \; \sum_{i=1}^{p} |d_i - \max\{c_i, q_i(x)\}| \\ &\text{subject to: } s \in \mathcal{B}(0, \Delta_k) \end{aligned}$$

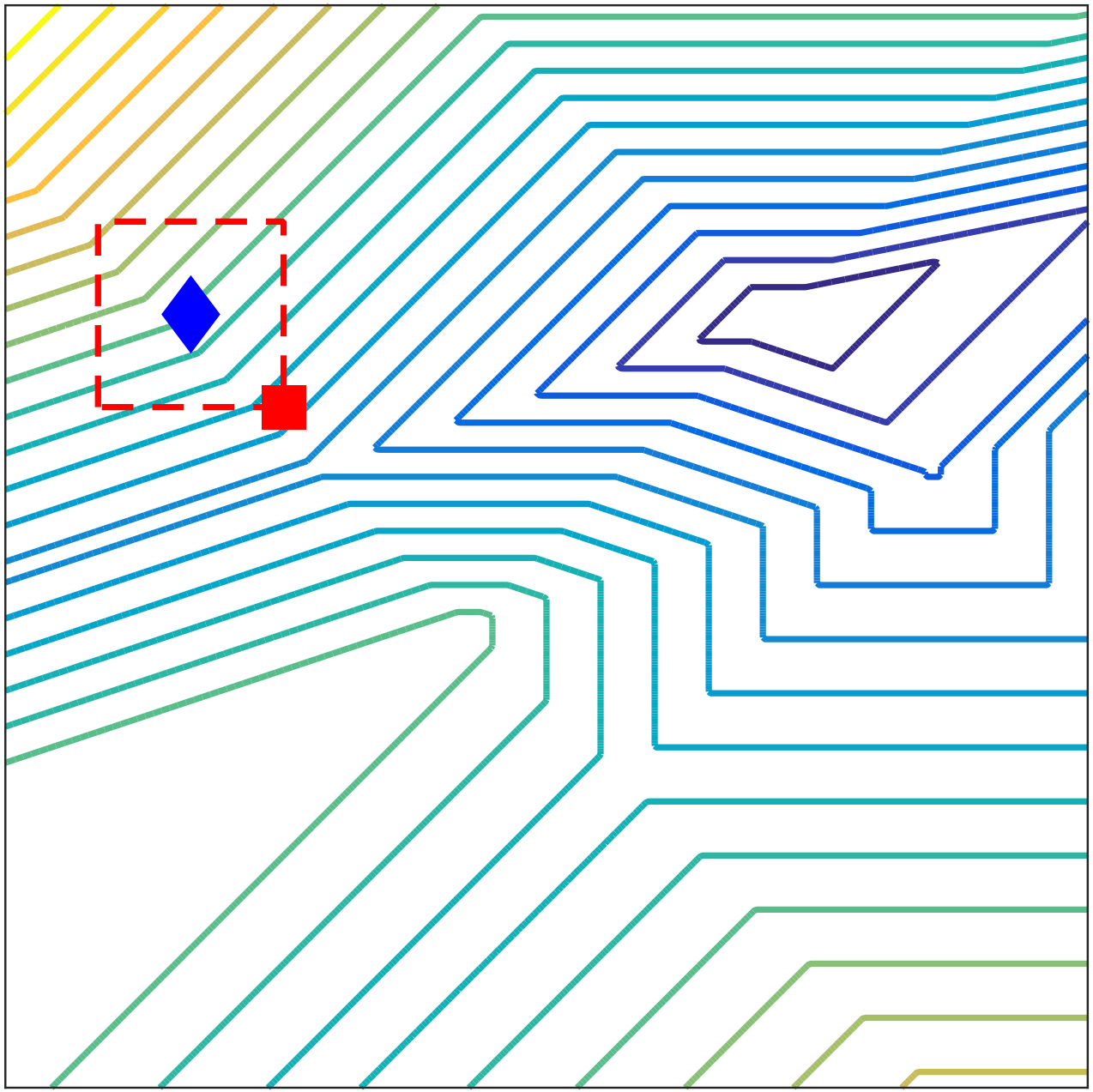
Manifold Sampling



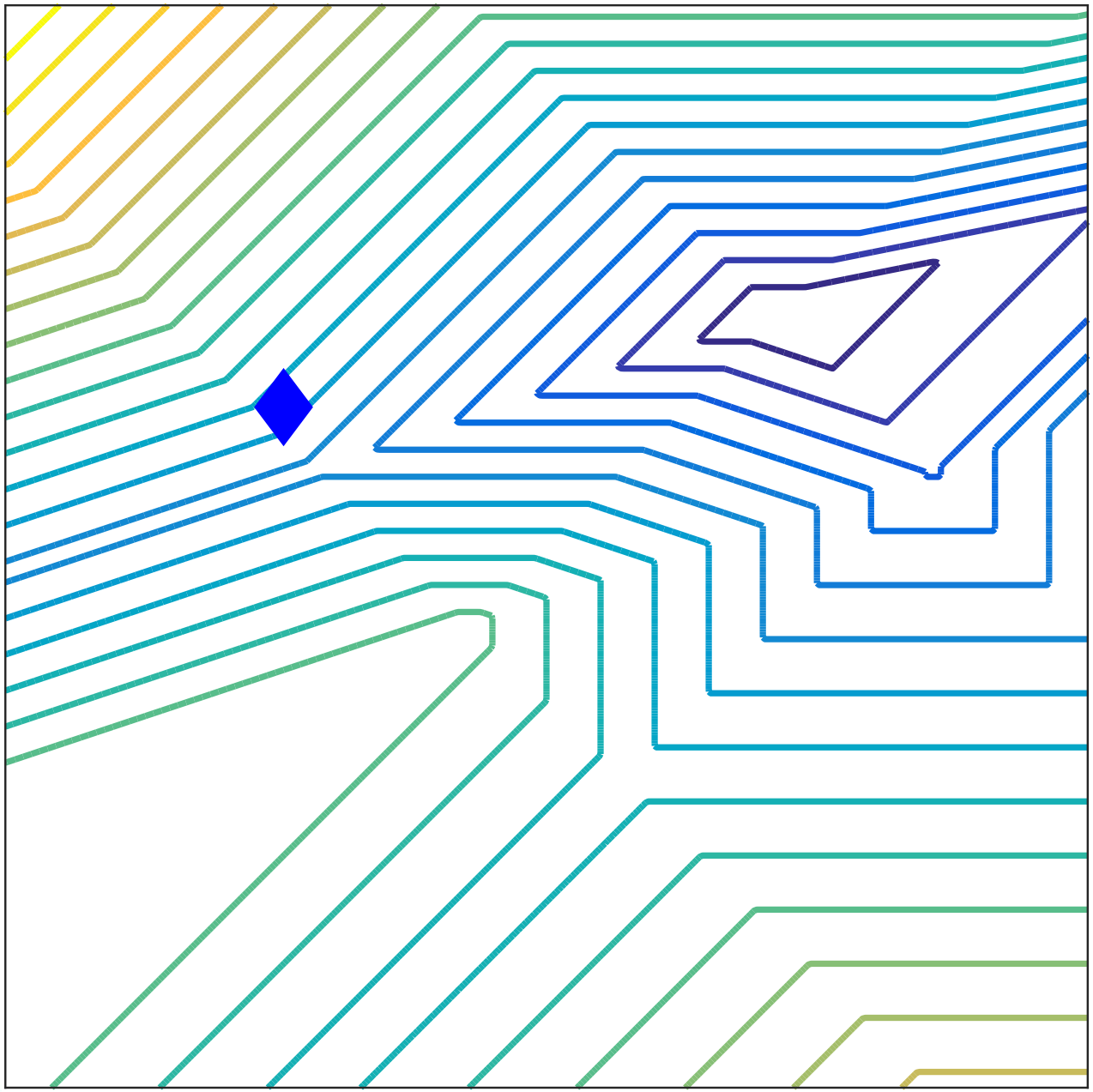
Manifold Sampling



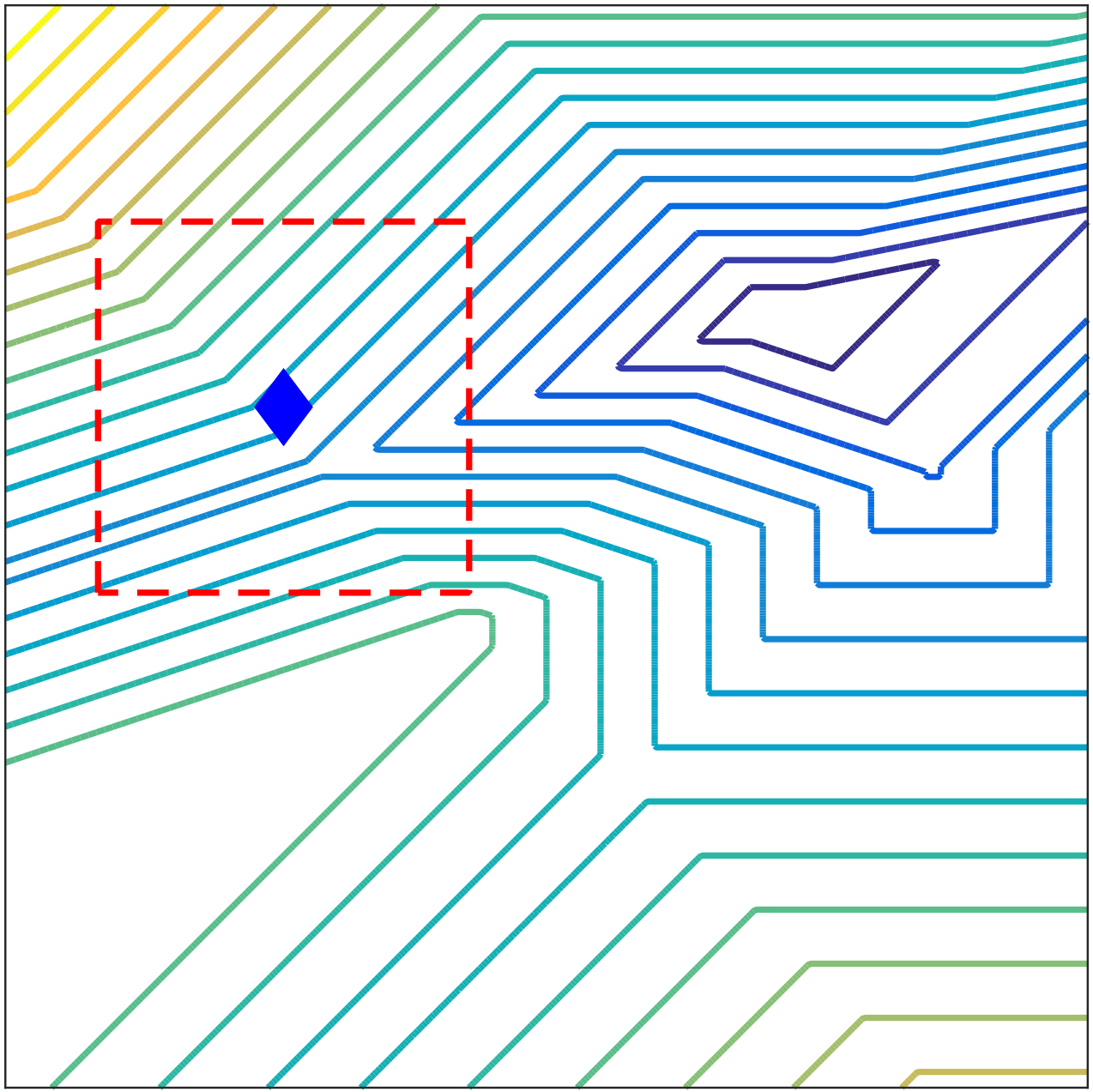
Manifold Sampling



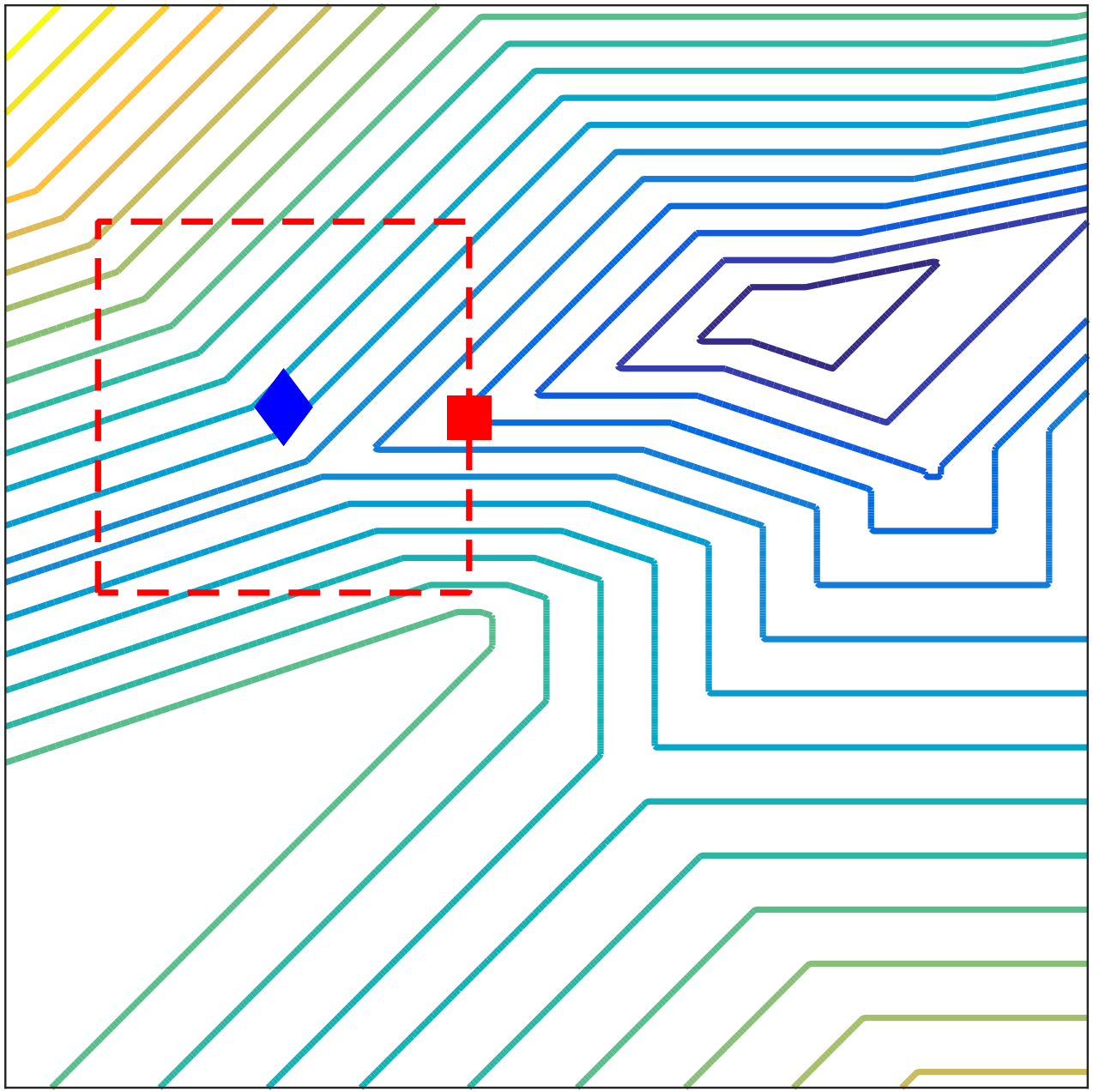
Manifold Sampling



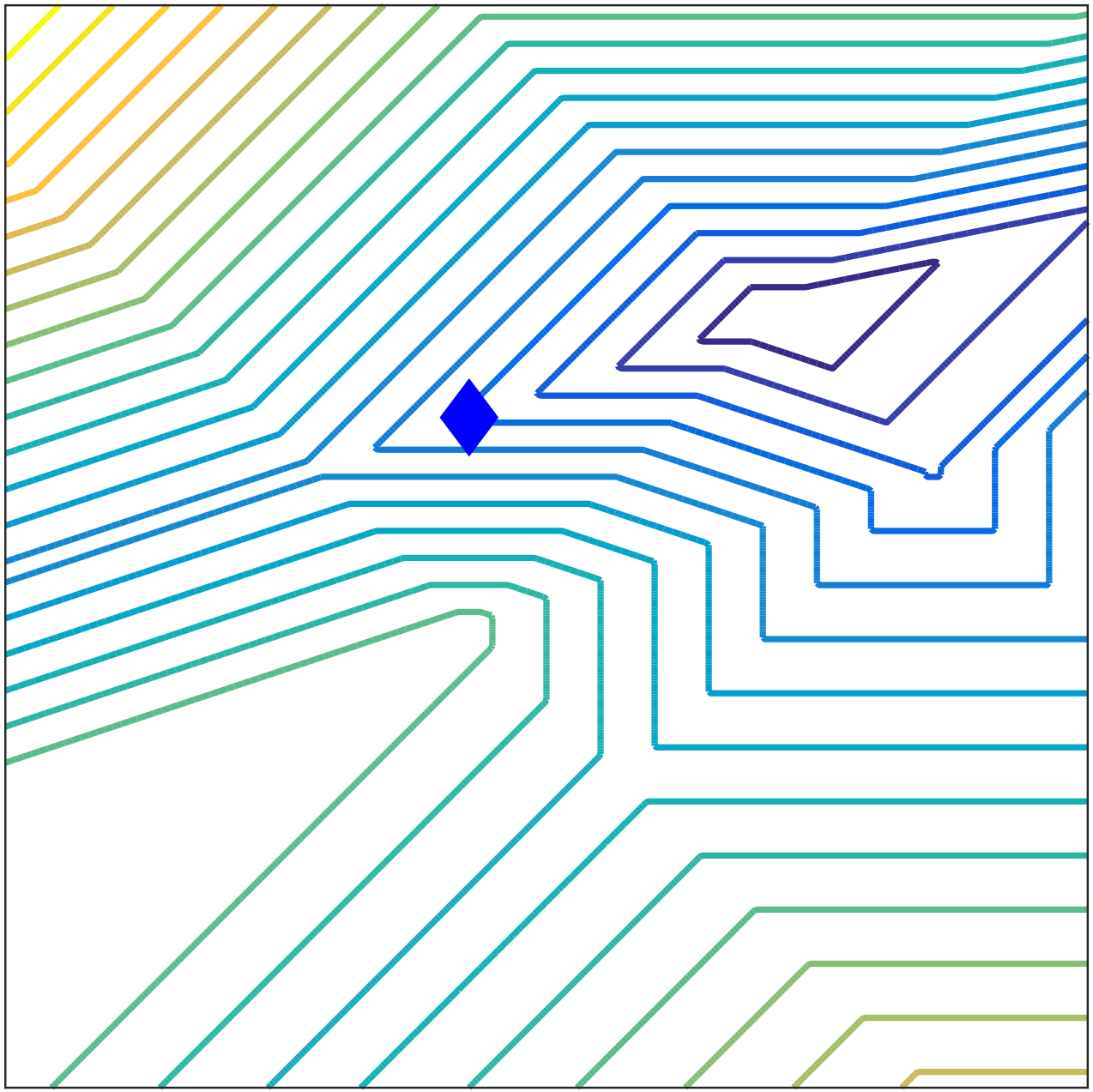
Manifold Sampling



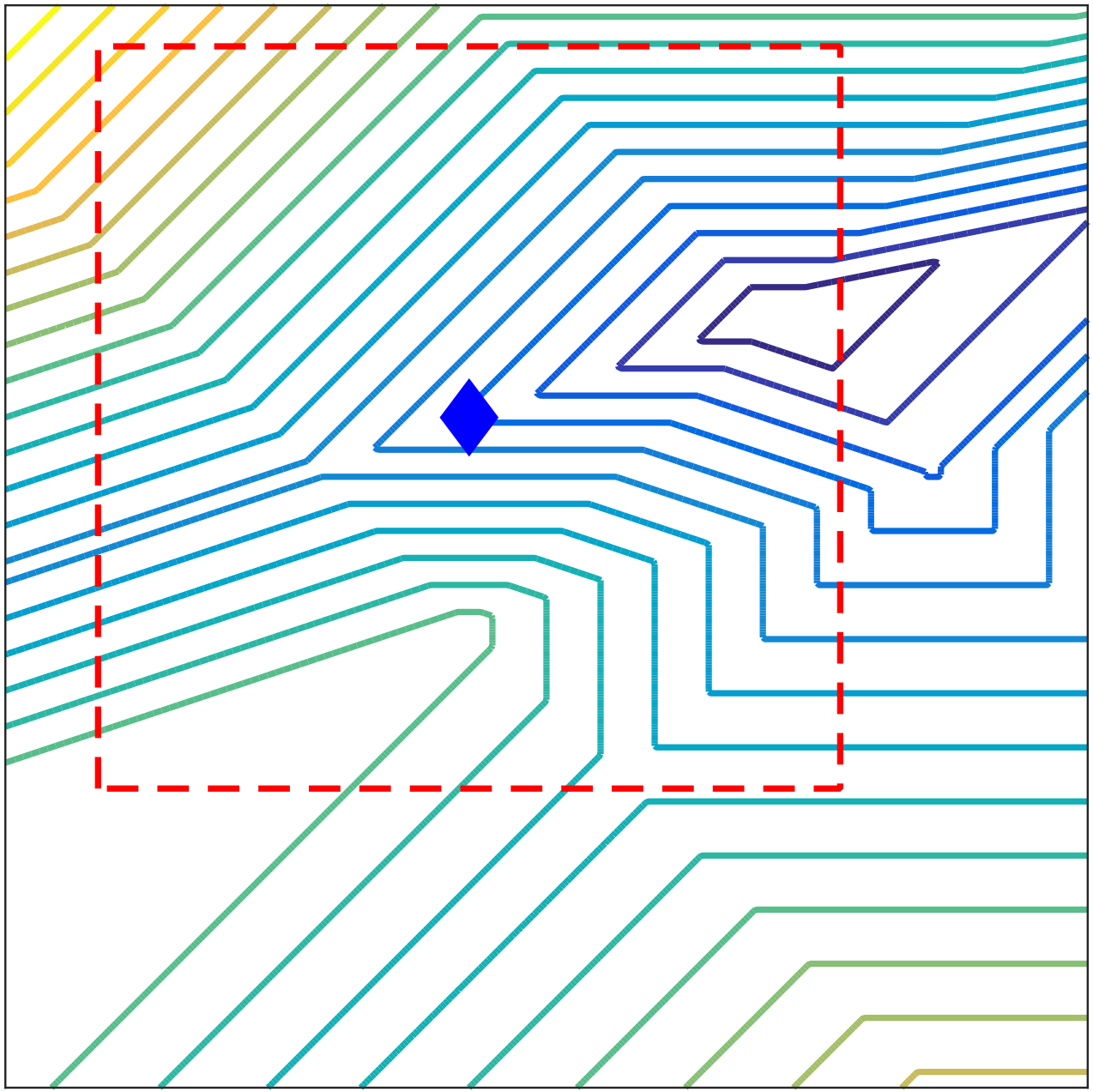
Manifold Sampling



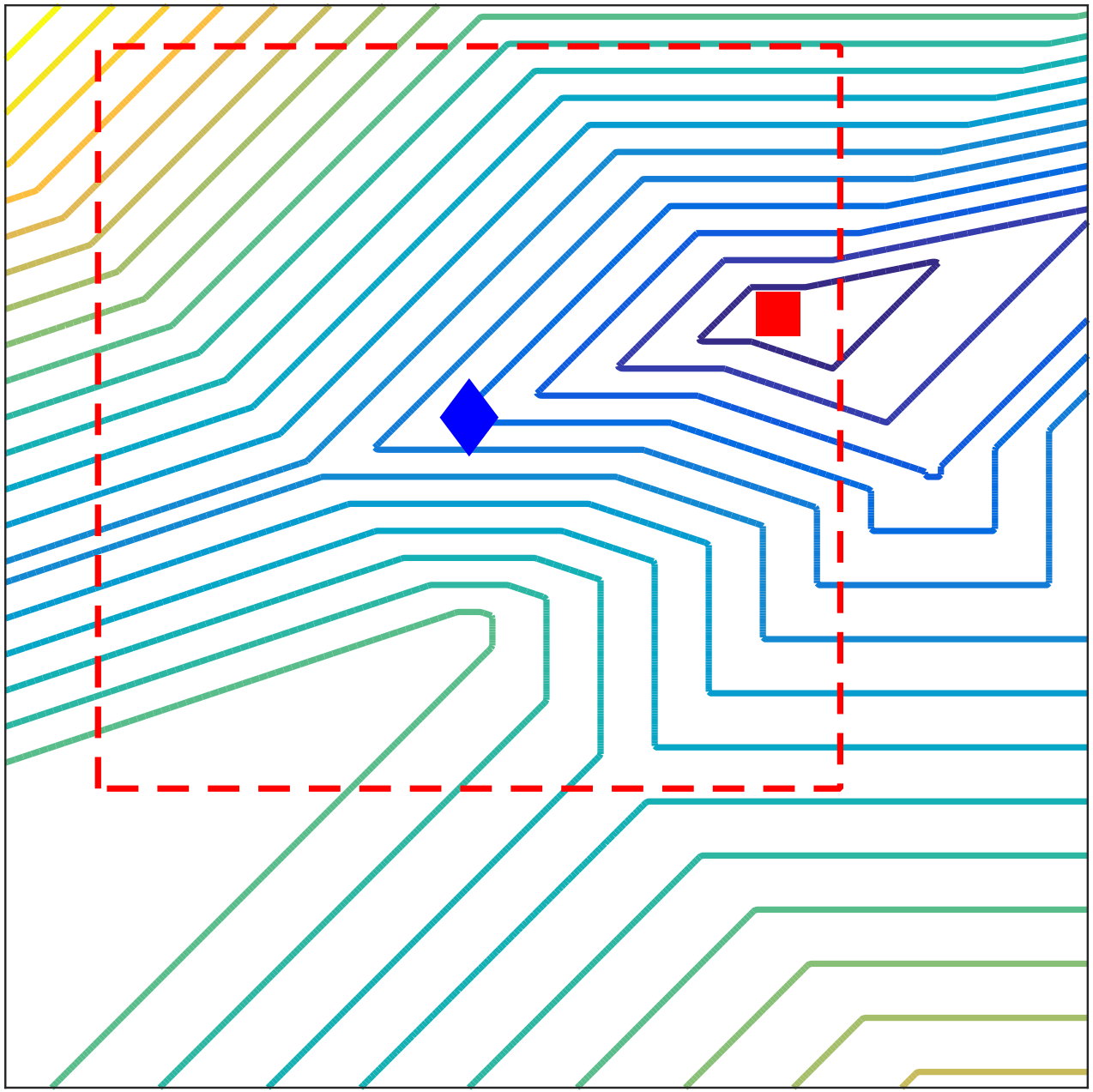
Manifold Sampling



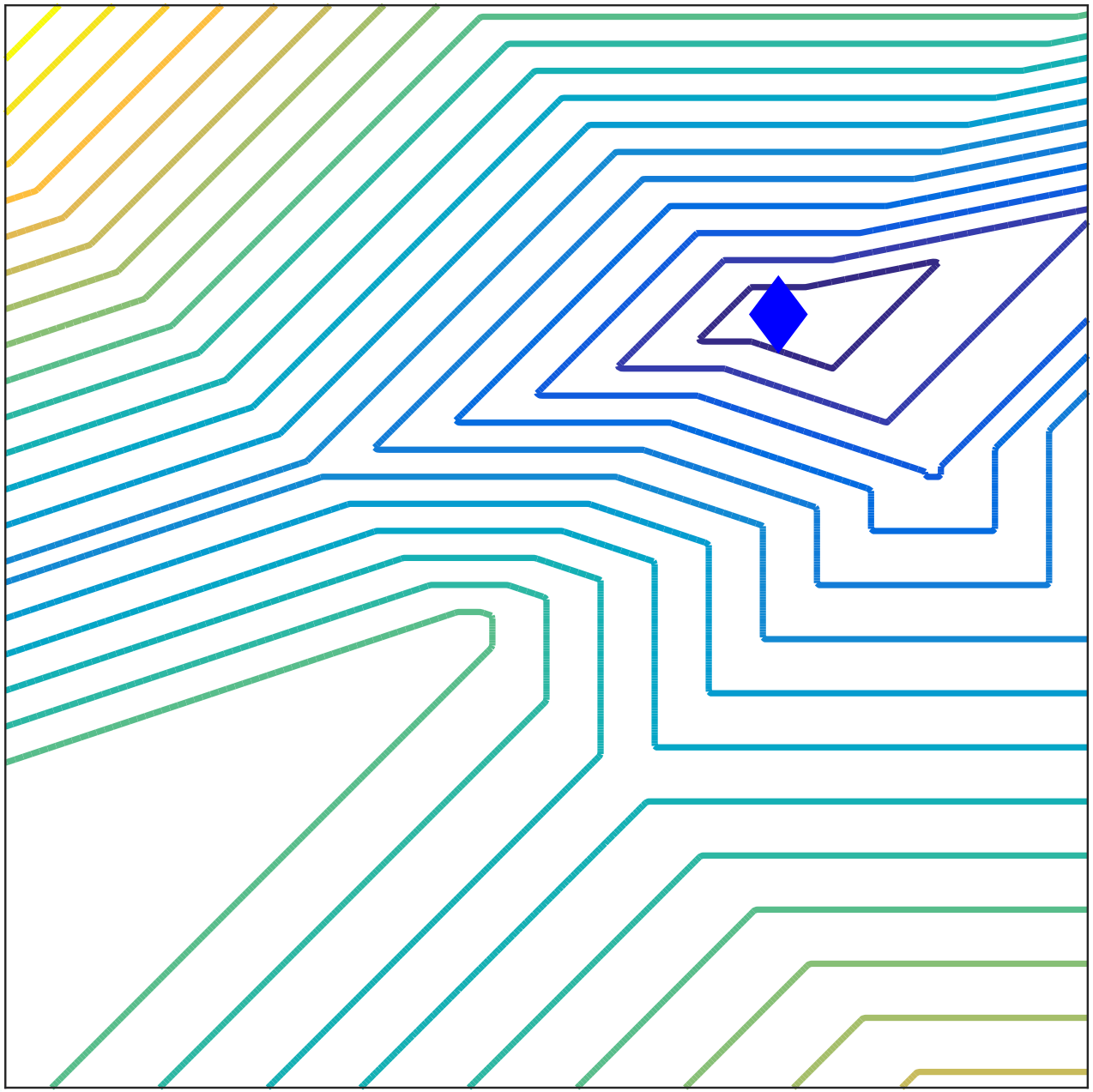
Manifold Sampling



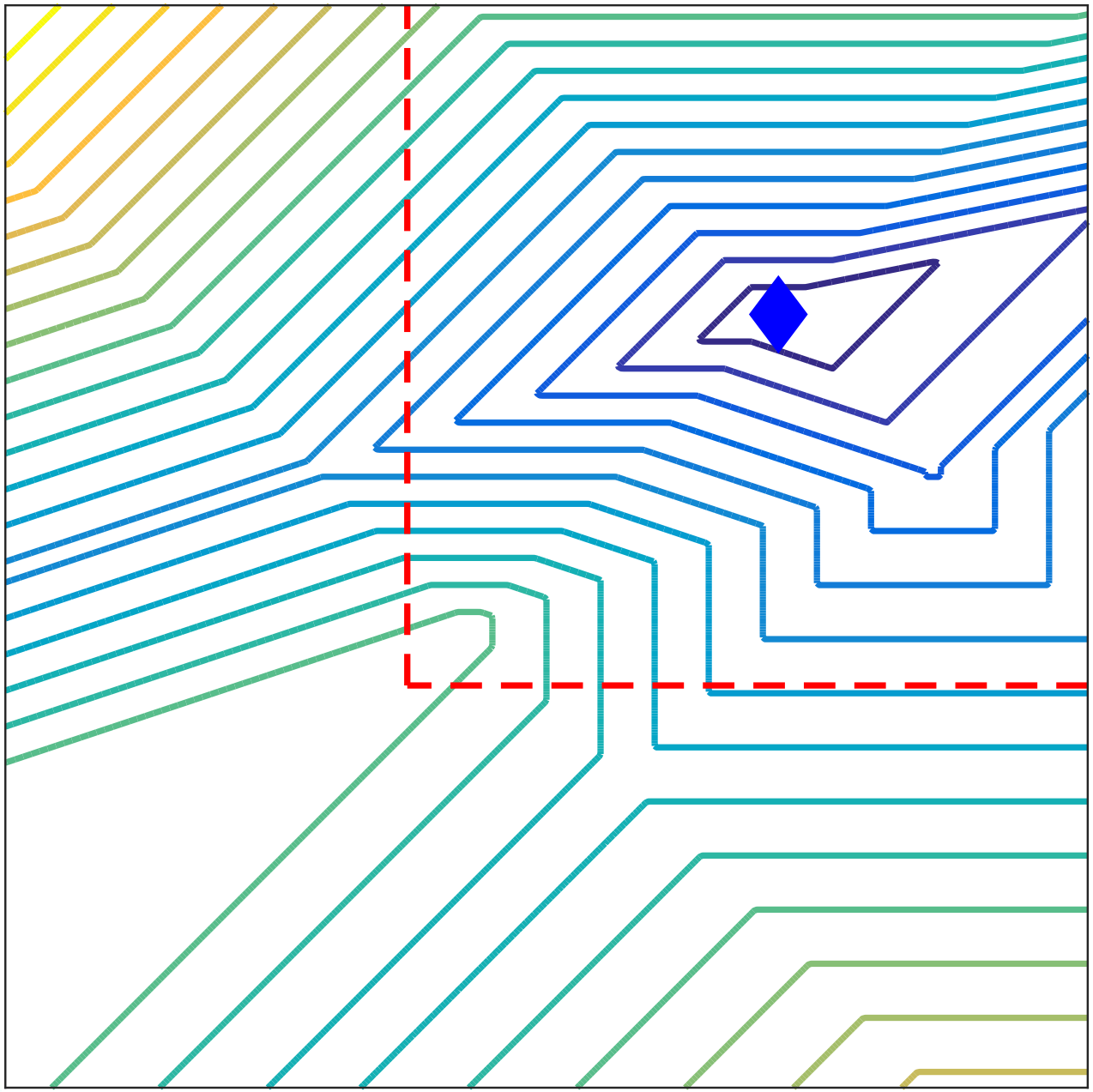
Manifold Sampling



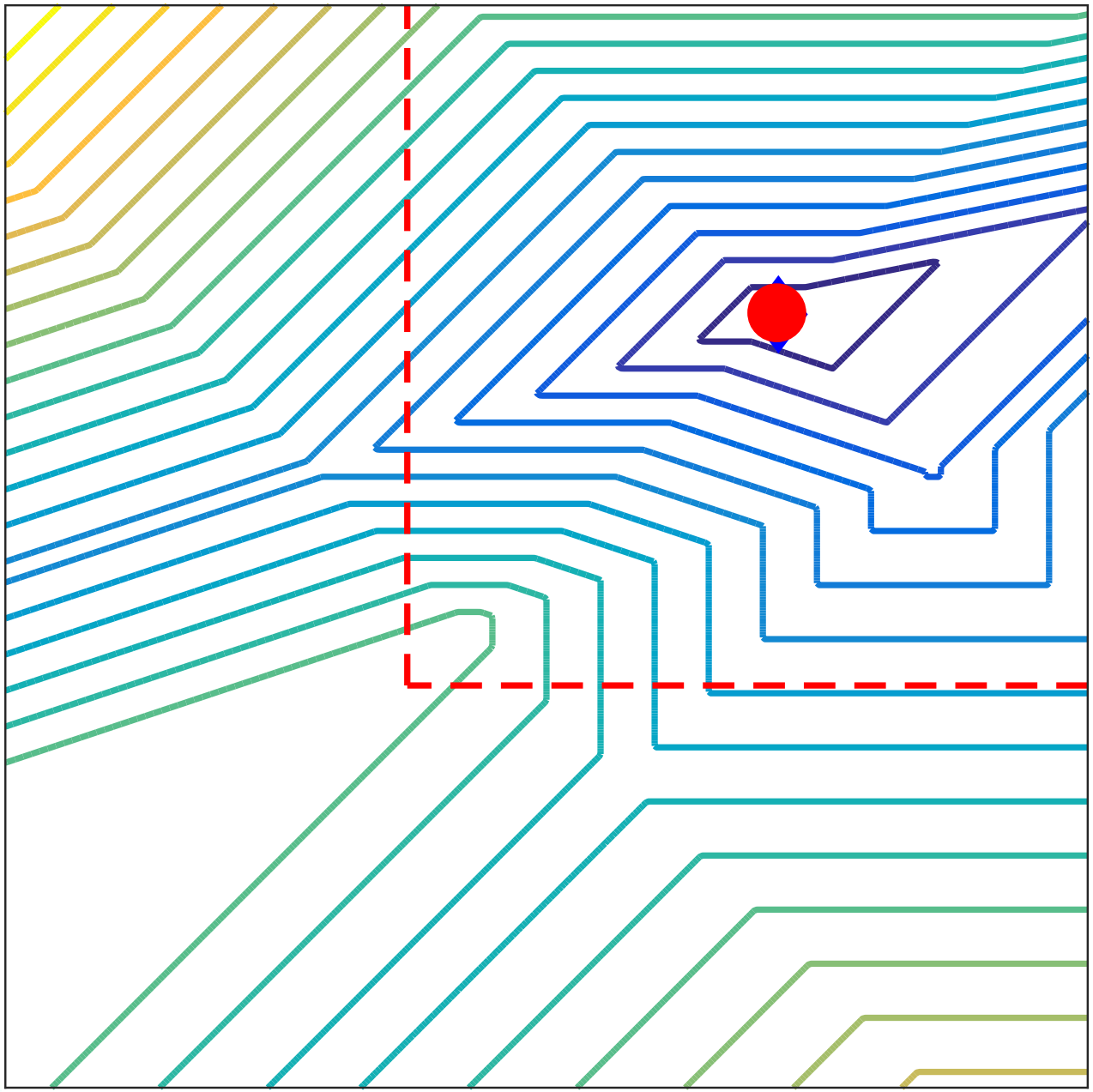
Manifold Sampling



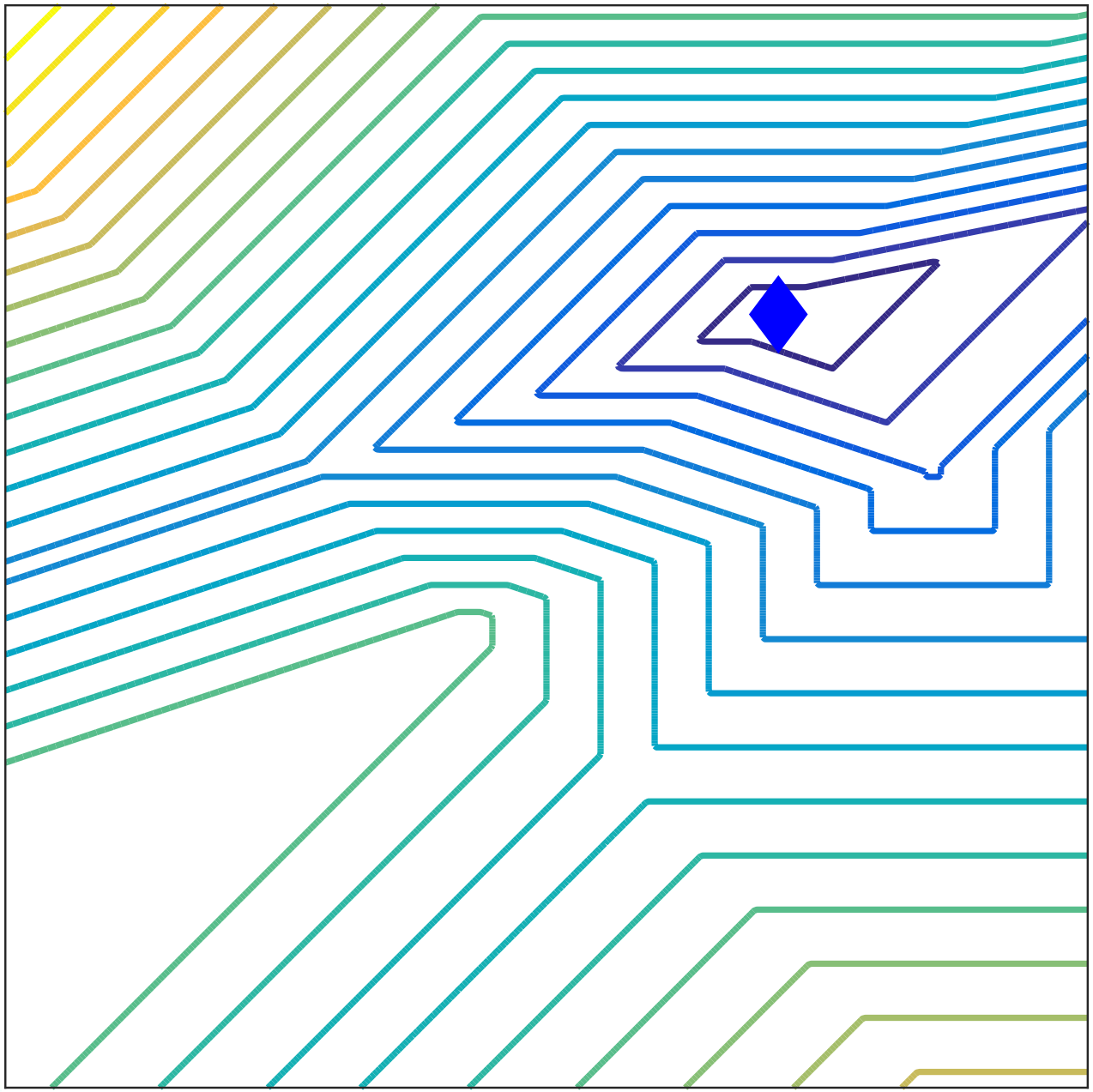
Manifold Sampling



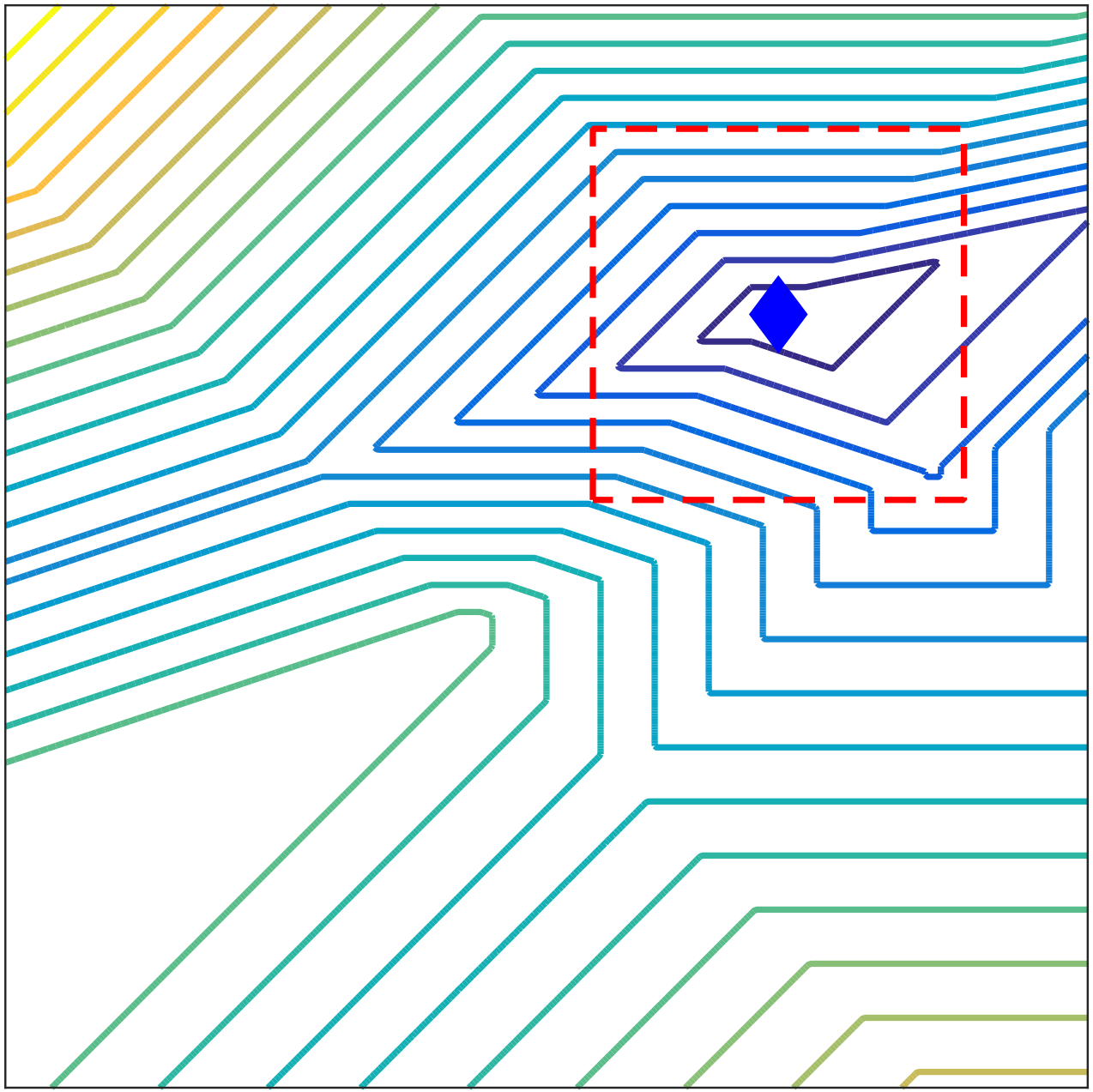
Manifold Sampling



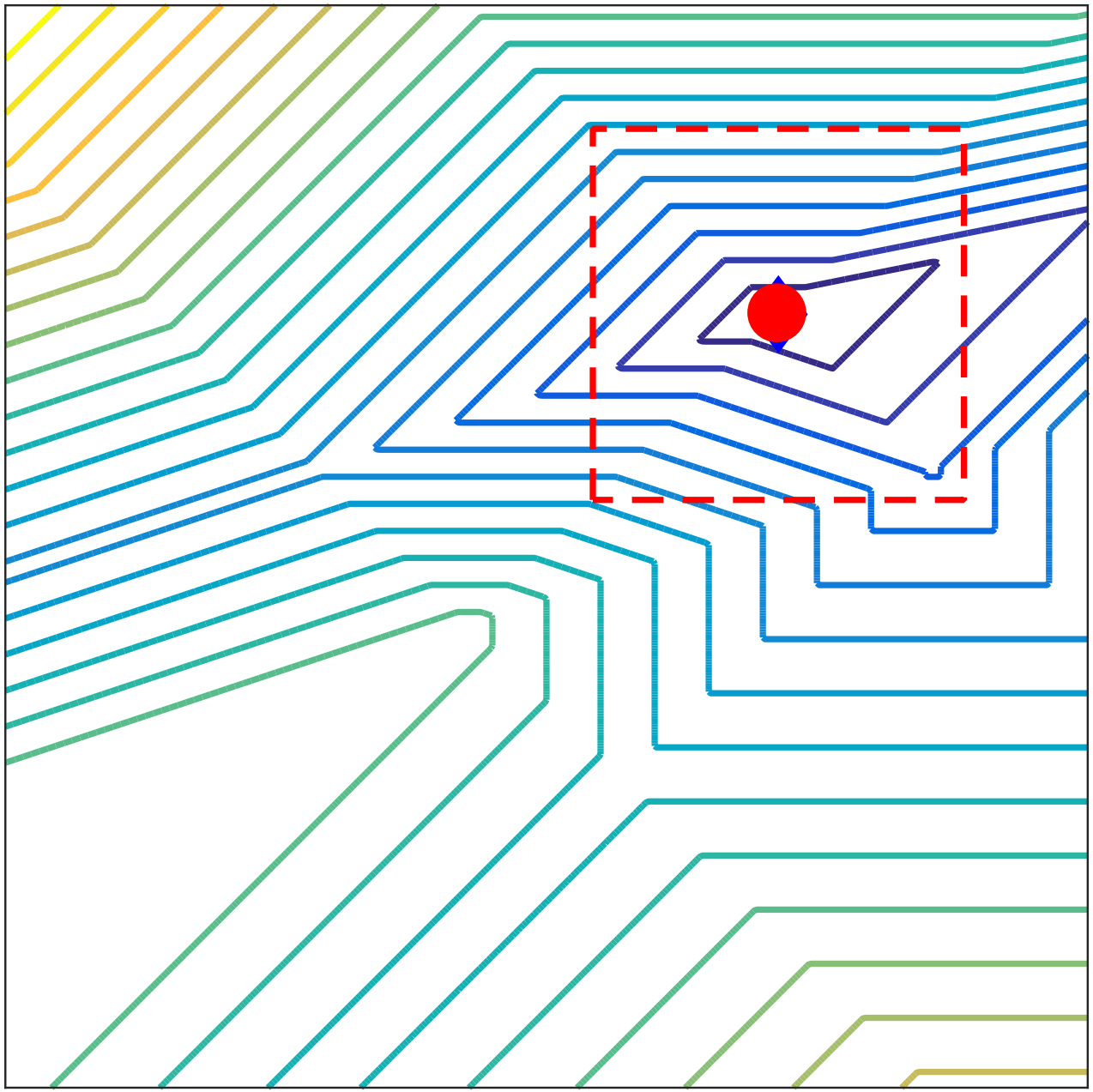
Manifold Sampling



Manifold Sampling



Manifold Sampling



Better trust-region subproblem?

Instead of solving

$$\begin{aligned} &\underset{s}{\text{minimize}} && m_k^f(x^k + s) \\ &\text{subject to:} && s \in \mathcal{B}(0, \Delta_k) \end{aligned}$$

How about

$$\begin{aligned} &\underset{s}{\text{minimize}} && h(M(x^k + s)) \\ &\text{subject to:} && s \in \mathcal{B}(0, \Delta_k) \end{aligned}$$

For censored ℓ_1 loss:

$$\begin{aligned} &\underset{s}{\text{minimize}} && \sum_{i=1}^{p} |d_i - \max\{c_i, q_i(x)\}| \\ &\text{subject to:} && s \in \mathcal{B}(0, \Delta_k) \end{aligned}$$

Better trust-region subproblem?

Instead of solving

$$\begin{aligned} &\underset{s}{\text{minimize}} \; m_k^f(x^k + s) \\ &\text{subject to: } s \in \mathcal{B}(0, \Delta_k) \end{aligned}$$

How about

$$\begin{aligned} &\underset{s}{\text{minimize}} \; h(M(x^k + s)) \\ &\text{subject to: } s \in \mathcal{B}(0, \Delta_k) \end{aligned}$$

For censored ℓ_1 loss:

$$\begin{aligned} &\underset{s}{\text{minimize}} \; \sum_{i=1}^{p} |d_i - \max\{c_i, q_i(x)\}| \\ &\text{subject to: } s \in \mathcal{B}(0, \Delta_k) \end{aligned}$$

Question

Best method for solving composite nonsmooth quadratic problems?

Measuring descent

- Descent is measured using a linearization $h^{(k)}$ of some selection function $\bar{h}$ and not h

Measuring descent

- Descent is measured using a linearization $h^{(k)}$ of some selection function $\bar{h}$ and not h

- Must ensure information about $\bar{h}$ is in $\mathbb{G}^k$ before taking a step

Measuring descent

- Descent is measured using a linearization $h^{(k)}$ of some selection function $\bar{h}$ and not h
- Must ensure information about $\bar{h}$ is in $\mathbb{G}^k$ before taking a step
- $h^{(k)}$ must satisfy

$$h^{(k)}(F(x^k)) \leq h(F(x^k)) \qquad \text{and} \qquad h^{(k)}(F(x^k+s^k)) \geq h(F(x^k+s^k)),$$

Measuring descent

- Descent is measured using a linearization $h^{(k)}$ of some selection function $\bar{h}$ and not h

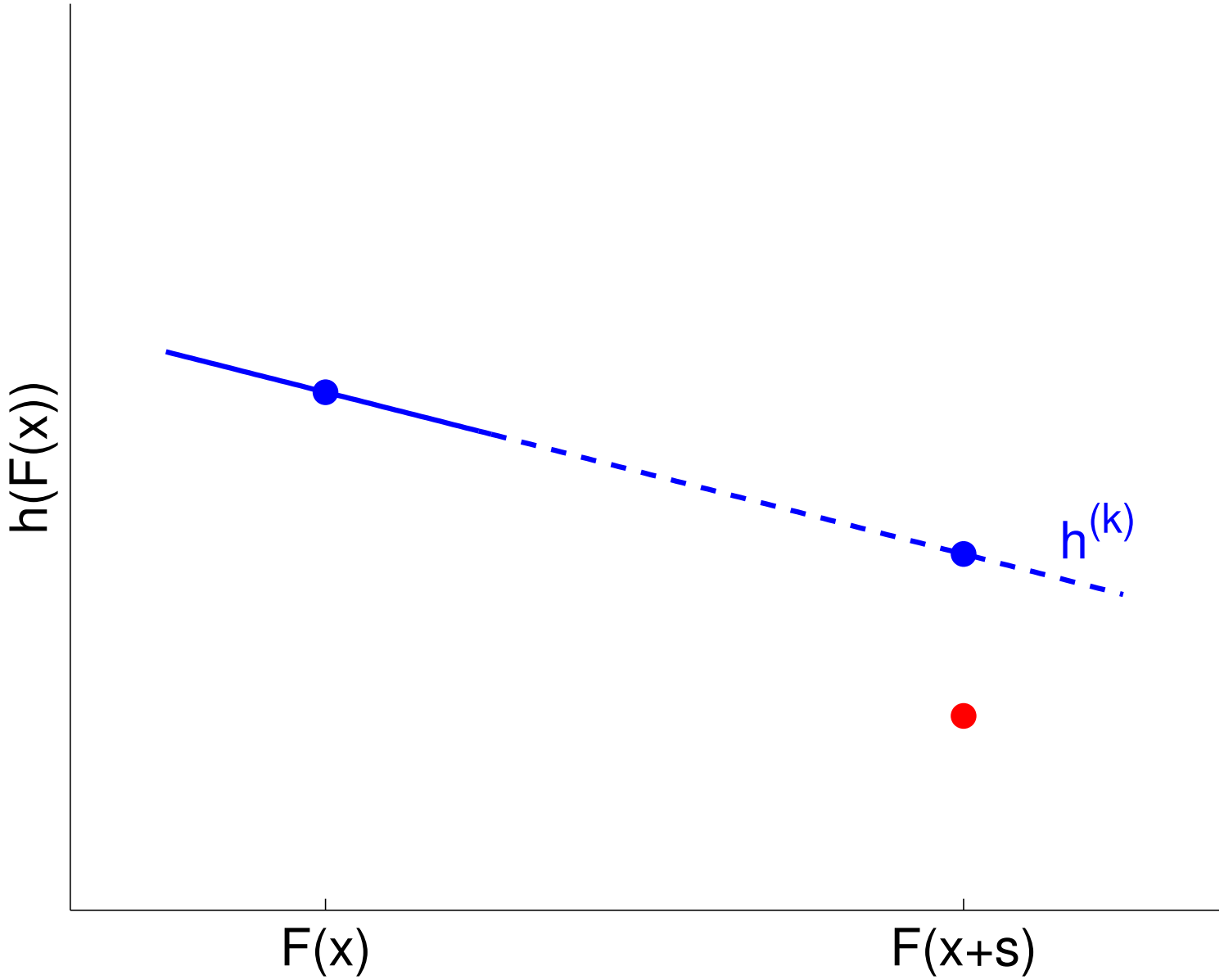
- Must ensure information about $\bar{h}$ is in $\mathbb{G}^k$ before taking a step

- $h^{(k)}$ must satisfy

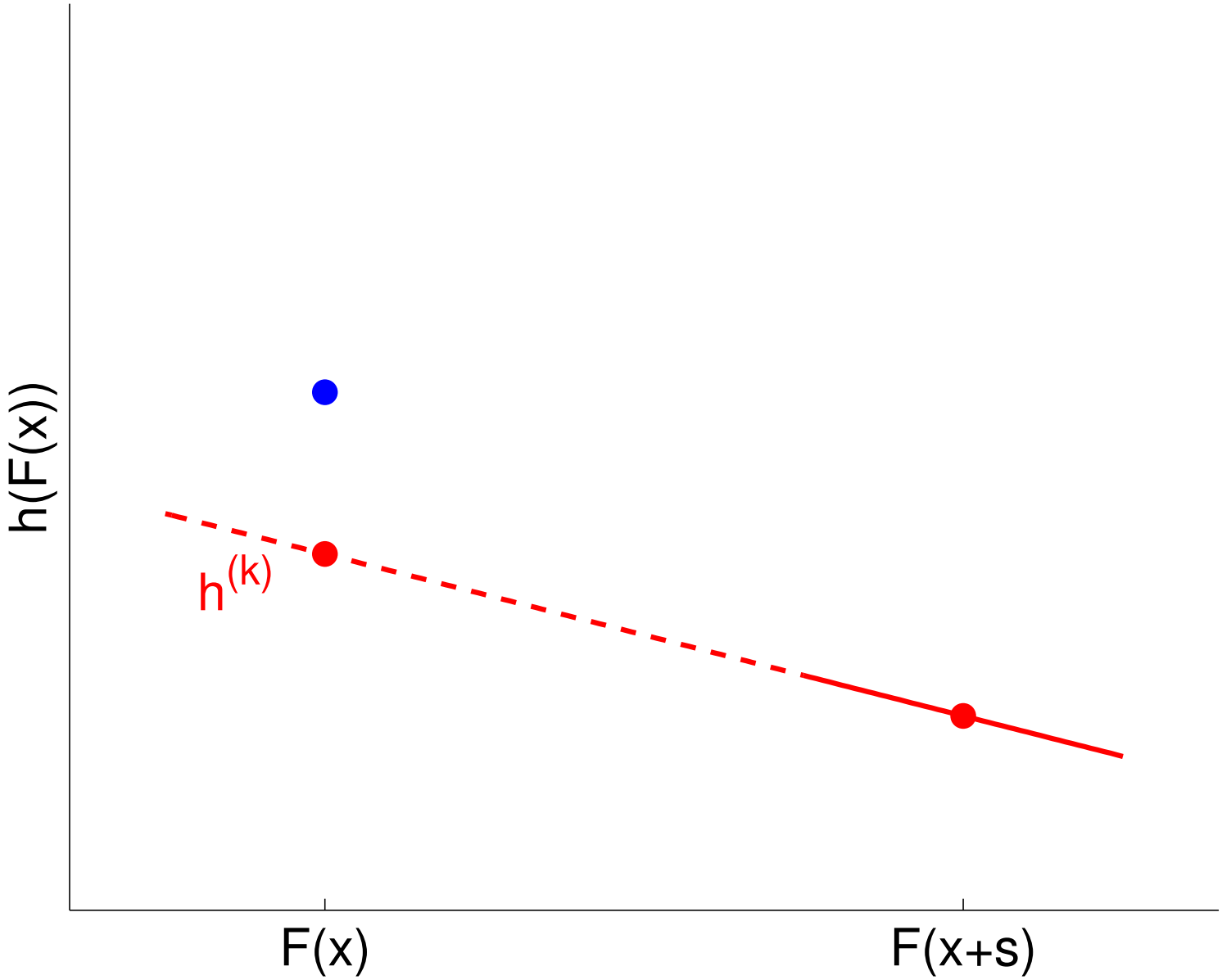
$$h^{(k)}(F(x^k)) \leq h(F(x^k)) \qquad \text{and} \qquad h^{(k)}(F(x^k+s^k)) \geq h(F(x^k+s^k)),$$

- $$\rho_k \triangleq \frac{h^{(k)}(F(x^k)) - h^{(k)}(F(x^k+s^k))}{m(x^k) - m(x^k+s^k)}$$

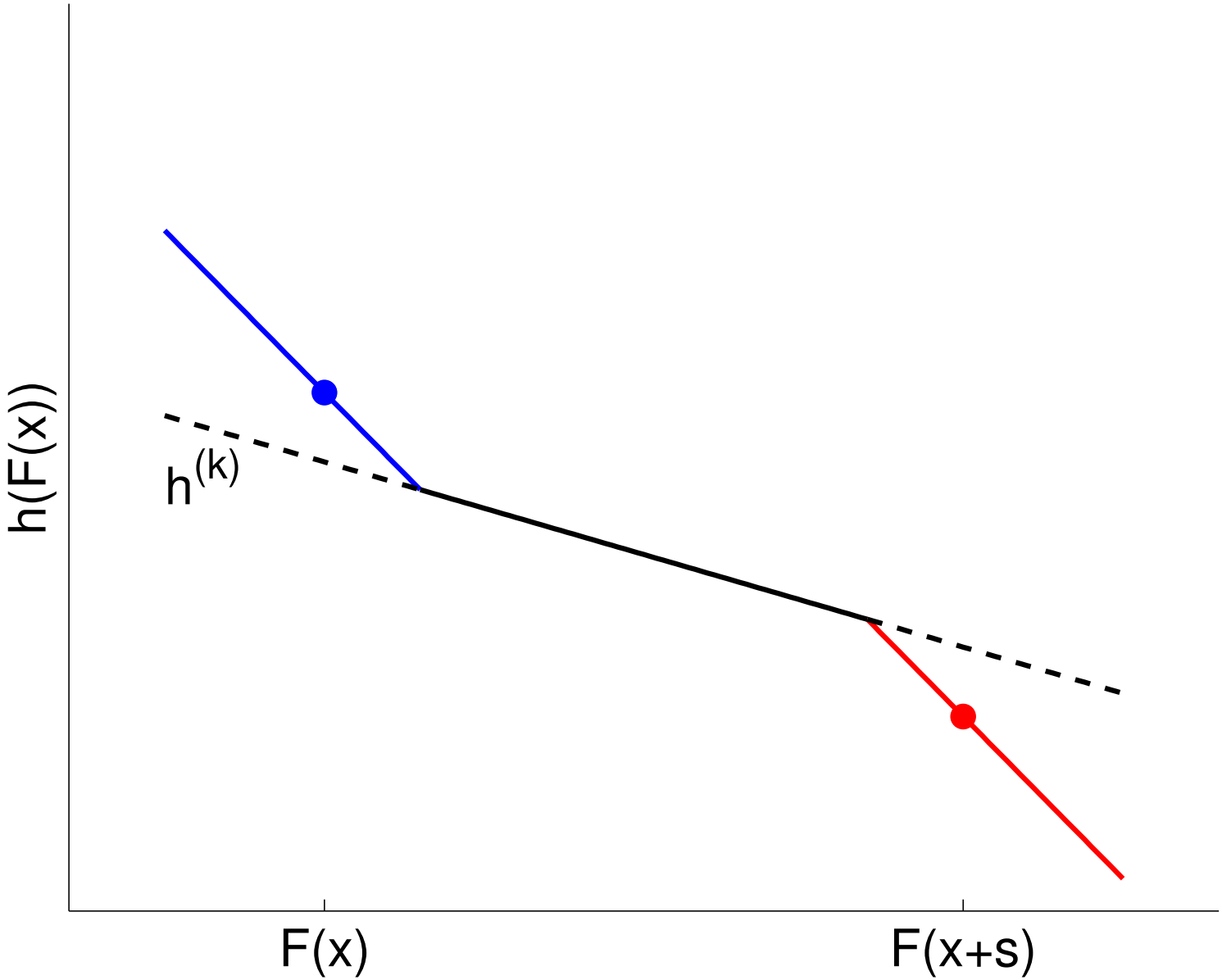
Examples of $h^{(k)}$



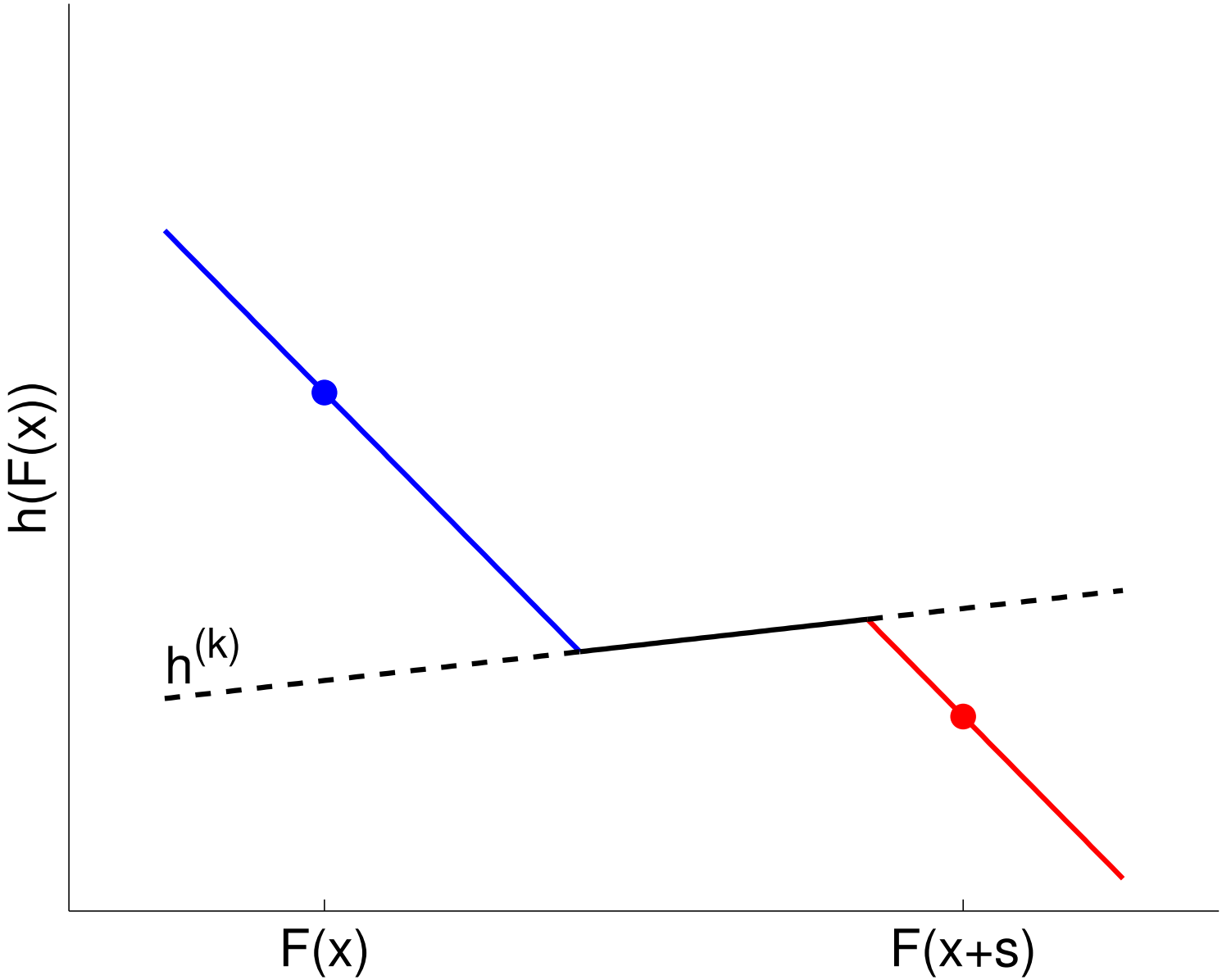
Examples of $h^{(k)}$



Examples of $h^{(k)}$



Examples of $h^{(k)}$



Convergence

- If the trust region radius Δ_k is a sufficiently small multiple of the model gradient $\left\|g^k\right\|$, the iteration is guaranteed to be successful.
- $\lim_{k\to\infty} \Delta_k = 0$.
- Some subsequence of master model gradients g^k goes zero.
- Zero is in the generalized Clarke subdifferential of cluster points of any subsequence of iterates with master model gradients converging to zero.
- The same holds for cluster points of the entire sequence of iterates.

Conclusions

When optimizing functions of the form $h(F(x))$ when

- h is "easy"
- F is "hard"

it can be advantageous to model F_i and then combine those models via known information about h.

Conclusions

When optimizing functions of the form $h(F(x))$ when

- h is "easy"
- F is "hard"

it can be advantageous to model F_i and then combine those models via known information about h.

jmlarson@anl.gov

Thank you!